University of Alberta

Evolutionary and functional characterization of genes in the cat eye syndrome critical region

by

Lindsay Jane Bridgland

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of Master of Science

in

Molecular Biology and Genetics

Department of Biological Sciences

Edmonton, Alberta
Spring 2003

National Library of Canada
Acquisitions and Bibliographic Services
395 Wellington Street
Ottawa ON K1A 0N4
Canada

Bibliothèque nationale du Canada
Acquisitions et services bibliographiques
395, rue Wellington
Ottawa ON K1A 0N4
Canada

Your file Votre référence

Our file Notre référence

The author has granted a non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of this thesis in microform, paper or electronic formats.

L'auteur a accordé une licence non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de cette thèse sous la forme de microfiche/film, de reproduction sur papier ou sur format électronique.

The author retains ownership of the copyright in this thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without the author's permission.

L'auteur conserve la propriété du droit d'auteur qui protège cette thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

0-612-82252-4

Canada

University of Alberta

Library Release Form

Name of Author: Lindsay Jane Bridgland

Title of Thesis: Evolutionary and functional characterization of genes in the cat eye syndrome critical region

Degree: Master of Science

Year this Degree Granted: 2003

Permission is hereby granted to the University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly, or scientific research purposes only.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

L Bridgland

Lindsay Jane Bridgland
9838 79th Avenue NW
Edmonton, Alberta
Canada T6E 1R1

Jan 23/03

University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled "Evolutionary and functional characterization of genes in the cat eye syndrome critical region" submitted by Lindsay Jane Bridgland in partial fulfillment of the requirements for the degree of Master of Science in Molecular Biology and Genetics.

Heather McDermid
Dr. H. E. McDermid

Warren Gallin
Dr. W. Gallin

Dr. M. Walter

Jan 2003

Abstract

Cat eye syndrome (CES) is a developmental disorder resulting from a duplication of chromosome 22q11.2. Fourteen genes have been identified in the cat eye syndrome critical region (CESCR) to date. The proximal half of the CESCR is rich in duplications, whereas the distal portion is gene-rich. This research focused on characterizing *CECR7*, which is located in the proximal portion of the CESCR. *CECR7* was formed from three duplicons in the chromosome 22q pericentromere. *CECR7* exons show similarity to sequences on chromosomes 2, 5, 7, 10, 11, 12, 13, 14, 15, 16, 18, 19, 21, and elsewhere on 22. Based on PCR analysis of duplicon boundaries in various primate species, *CECR7* was formed before the separation of macaque. However, RT-PCR results indicate that expression didn't result immediately from the formation of this novel transcription unit, or that expression was silenced in orangutan following its formation. Characterization of *CECR1* ADA activity, a candidate gene for involvement in CES, was inconclusive.

Acknowledgements

I would like to thank Heather McDermid for her support over the last few years, as well as her encouragement and ideas. I am also grateful to my committee, comprised of Warren Gallin and Mike Walter, for their suggestions. I would like to thank Dr. Gallin for his assistance with the expression of the human ADA gene in mammalian cells.

I extend a special thanks to Graham Banting and Kathy Bueble for their friendship, and to Graham for the technical suggestions and assistance he provided throughout the years. As well, I would like to extend my sincere thanks to my family for encouraging me from afar and helping me to get to where I am today.

Table of Contents

Chapter 1: Introduction 1
1. 1 Chromosome 22 1
1.2 Chromosomal Disorders Related to Chromosome 22q11 1
1.3 Cat eye syndrome 3
1.4 Other Disorders with CES-like Phenotype 6
1.5 The Cat Eye Syndrome Critical Region (CESCR) 8
Chapter 2: Introduction to *CECR7* and *CECR8* 12
2.1 The Centromere 12
2.2 Alpha Satellite Repeats 12
2.3 The Pericentromeric Region 13
2.4 Mechanisms of Genome Evolution 14
2.5 Examples of Pericentromeric Duplications 16
2.5.1 NF1 16
2.5.2 Chromosome 10 18
2.5.3 16p11.1 Duplication 18
2.6 Examples of Novel Genes Created via Duplicons 22
2.6.1 ASMTL 22
2.6.2 The Drosophila Gene jingwei 23
2.6.3 CT2- Creatine Transporter 2 24
2.7 Research Objectives 26
Chapter 3: *CECR7* and *CECR8* Materials and Methods 27
3.1 Computer Analysis: Identification and Analysis of *CECR7* and *CECR8* 27
3.2 Primer Design 27
3.3 PCR: Cloning of Genes and Examination of the Presence of Duplicon Boundaries in Primates 28
3.4 Identification of Paralogous Locations of *CECR7*-like Sequence: PCR using Monochromosomal Hybrid Panel DNA 28
3.5 Reverse Transcription PCR (RT-PCR): Expression Analysis of *CECR7* and *CECR8* 29
3.6 Rapid Amplification of cDNA Ends (5' and 3' RACE): Cloning of *CECR7* and *CECR8* 29
3.7 Generation of a Testis cDNA Library 29
3.8 Primer Extension: Identification of the Transcription Start Site of *CECR7* 30
3.9 Northern Blot Transfer: Expression Analysis of *CECR7* and *CECR8* 31
3.10 Northern Hybridization 31
3.11 DNA Purification from Agarose Gels 32
3.12 Plasmid DNA Preparation and Purification 32
3.13 Sequencing 32
3.14 Tissue Culture: Growth of Primate Lymphoblast and Fibroblast Cells 33
3.15 Preparation of Freezer Stocks of Cell Lines 33
3.16 Isolation of DNA from Cell Lines 34
3.17 Isolation of RNA from Cell Lines 34
Chapter 4: Results 37
4.1 Cloning of *CECR7* 37
4.2 Expression Analysis of *CECR7* 43
4.3 Identification of *CECR7*-like Sequences Throughout the Human Genome 44
4.6 Alternative Splicing Identified in *CECR8* 67
4.7 Identification of *CECR8*-like sequence 68
4.8 Expression Analysis of *CECR8* 71

5.1 Significance of Proposed Origins of *CECR7* 73
5.3 Expression Analysis of *CECR7* 88
5.4 *CECR7*: A Putative Gene 91
5.5 *CECR8*: Significance of Proposed Origin 93
5.6 Expression Analysis of *CECR8* 94
5.7 Significance of Research 95
Chapter 6: Introduction to *CECR1* **99**
6.1 *CECR1* 99
6.2 Adenosine Deaminase (ADA) 101
6.3 Research Objectives 105
Chapter 7: *CECR1* Materials and Methods **106**
7.1 Creation of *CECR1* Constructs 106
7.2 Transfection of Various Insect Cell Lines 108
7.3 Generation of Stable Cell Lines Expressing *CECR1-A*, ADA-1, and ADA-2 109
7.4 Isolation of *CECR1-A*, ADA-1, and ADA-2 Protein 109
7.5 Analysis of Protein Expression: SDS-PAGE 110
7.6 Western Blot Analysis 111
7.7 Adenosine Deaminase Assay 111
Chapter 8: *CECR1* Results **115**
8.1 Expression of Recombinant Proteins and Analysis of ADA Activity 115
Chapter 9: *CECR1* Discussion **125**
9.1 Expected Results versus Apparent Results: *CECR1-A* 125
9.2 Future Directions 128
9.3 Significance of Research 130
Chapter 10: Conclusion **131**
10.1 Overall Significance of Research: A Final Summary 131
References **134**

List of Tables

Table 1-1: Summary of the Percentage of Patients Presenting with CES Phenotype..... 11

Table 3-1: List of Available *CECR7* Primers..... 36

Table 4-1: PCR Primers Used in the Cloning and Characterization of *CECR7* and *CECR8*..... 38

Table 7-1: PCR Primers Used in Creating Constructs for Expression in Insect Cells..... 113

Table 7-2: PCR Primers Used to Sequence Constructs..... 114

Table 8-1: Assay Results Using Commercial ADA..... 122

Table 8-2: ADA Assay Results Using a Dilution Series of Commercial ADA..... 123

Table 8-3: ADA Assay Results for the Insect Expressed Proteins..... 124

List of Figures

Figure 1: The cat eye syndrome marker chromosome (CEC) 4
Figure 1-2: A map of the cat eye syndrome critical region 9
Figure 2-1: Summary of duplication events that occurred between chromosome 4, X, and 16 21
Figure 4-1: Genomic structure of *CECR7* 40
Figure 4-2: Exon 1 RT-PCR results 42
Figure 4-3: Monochromosomal hybrid panel PCR results spanning *CECR7* intron 1 and intron 3 duplicon boundaries 46
Figure 4-4: NIGMS panel #2 47
Figure 4-5: Similarity of *CECR1* to hypothetical gene 50
Figure 4-6: Summary diagram illustrating the complex nature of the region surrounding *CECR7* 51
Figure 4-7: Similarity of *CECR7* to the two coenzyme A genes on chromosome 16 .53
Figure 4-8: Similarity of *CECR7* exons 4, 4', and 4'' to chromosome 12 55
Figure 4-9: Verified structure of chromosome 12 transcript 56
Figure 4-11: PCR results spanning the intron 3 duplicon boundary 59
Figure 4-12: Alignment of human and primate intron 1 boundary sequences 60
Figure 4-13: Alignment of human and primate intron 3 boundary sequences 61
Figure 4-14: RT-PCR results indicated *CECR7* is expressed in gorilla fibroblasts 64
Figure 4-15: RT-PCR results demonstrating *CECR7* expression in human and gorilla fibroblasts, but not orangutan or macaque 65
Figure 4-16: Alignment of *CECR7* RT-PCR products from human and gorilla 66
Figure 4-17: Alternative splicing in *CECR8* 69
Figure 4-18: Similarity of *CECR8* to sequence on the Y chromosome 70
Figure 4-19: *CECR8* Northern blot results 72
Figure 5-1: Hypothesized duplication events leading to the presence of *CECR7*-like sequence on chromosomes 22, 12p13.3, and 12p12 79
Figure 5-2: The primate lineage 84
Figure 6-1: Schematic illustration of the coding region of *CECR1* 100
Figure 7-1: Comparison of the MCS of pMIB/V5-HIS A and the modified vector..107
Figure 8-1: Comparison of the amino acid sequences of ADA*2 and *CECR1* 116
Figure 8-2: ADA-2 RT-PCR Results 117
Figure 8-3: Western Blot Results for Insect Expressed Proteins 120

Abbreviations

ABI: applied biosystems incorporated
ADA: adenosine deaminase
ASMTL: N-acetylserotonin O-methyltransferase-like
bp: base pair
BAC: bacterial artificial chromosome
BUCS1: butyryl coenzyme A synthetase 1
CCRI: Coriell Cell Repository
CEC: cat eye chromosome
CECR1: cat eye syndrome critical region gene 1
CECR7: cat eye syndrome critical region gene 7
CECR8: cat eye syndrome critical region gene 8
CENP-B: centromere-specific DNA binding protein B
CES: cat eye syndrome
CESCR: cat eye syndrome critical region
CHARGE: charge association
CT: creatine transporter
CTAB: cethyltrimethylammonium bromide
DGS: Digeorge Syndrome
DNTPs: dinucleotide triphosphates
DTAB: dodecyltrimethylammonium Bromide
ECACC: European Collection of Cell Cultures
EST: expressed sequence tag
FBS: fetal bovine serum
FISH: fluorescent in situ hybridization
htgs database: high-throughput genomic sequence
kb: kilobase
kDA: kilodaltons
LCR: low copy repeats
Mb: megabase
mA: milli-amps
mya: million years ago
NF1: neurofibromatosis 1
nr database: non-redundant database
ORF: open reading frame
PAC: P1 artificial chromosome
PAR1: pseudoautosomal region 1
PCR: polymerase chain reaction
RACE: rapid amplification of cDNA ends
RPM: revolutions per minute
RT: reverse transcription
SHH: sonic hedgehog
TAPVR: total anomalous pulmonary venous return
TOF: tetralogy of fallot
VACTERL: vertebral anomalies, anal atresia, cardiac malformations, trascheoesophageal fistula, renal anomalies, limb anomalies
VCFS: Velo-Cardial Facial Syndrome
YAC: yeast artificial chromosome
ZNF: zinc finger

Chapter 1: Introduction

1. 1 Chromosome 22

Chromosome 22 is the second smallest human autosome. This acrocentric chromosome was the first chromosome to be fully sequenced (Dunham *et al.*, 1999). The sequence was determined by a consortium of sequencing centres including the Sanger Centre (www.sanger.ac.uk/HGP/Chr22), Keio University School of Medicine, the University of Oklahoma, and the Genome Sequencing Centre. Several gaps still exist in the sequence, but with current methodologies further sequencing of these regions is very difficult and progress in these regions has been slow.

The short arm of chromosome 22 (22p) is highly similar to the short arms of the other acrocentric chromosomes (13, 14, 15, and 21). Chromosome 22p contains a series of tandem repeat structures, some of which encode structural ribosomal RNAs. The protein coding genes are found on the long arm of chromosome 22 (22q) and several chromosomal anomalies leading to syndromes have been associated with 22q.

1.2 Chromosomal Disorders Related to Chromosome 22q11

Several genomic disorders are associated with chromosome 22q11 (reviewed in McDermid and Morrow 2002). The most common chromosomal abnormality associated with chromosome 22q11 is a deletion, which causes Velo-Cardio-Facial-Syndrome (VCFS) (MIM 192430), which was first described by Strong (1968). VCFS patients present with a variable phenotype, the most common features being cleft palate, learning disabilities, cardiac malformations, typical facial features, and ear anomalies (Arnold *et al.*, 2001). Schizophrenia and paranoid delusions have also been associated with VCFS (Arnold *et al.* 2001). The estimated incidence of VCFS in

the population is 1/4500, with 90% of cases occurring sporadically (Arnold *et al.*, 2001). A microdeletion in chromosome 22q11.2 has been shown to be the cause of this disorder. A second syndrome, DiGeorge Syndrome (DGS) (MIM 188400), also results from a submicroscopic deletion of chromosome 22q11.2. DiGeorge patients typically present with conotruncal heart defects, distinct facial features, hypoparathyroidism, and immune deficiency (Carey *et al.*, 1992; Swillen *et al.*, 2000). It is now clear that VCFS and DGS are actually caused by varying expressivity of the same microdeletion (Arnold *et al.*, 2001). However, what is not clear is whether this chromosome 22q11 deletion syndrome is a contiguous gene deletion syndrome or a monogenic disorder. Mouse models suggest that the only murine haploinsufficient gene in the DiGeorge region is *Tbx1* (Jerome *et al.* 2001; Lindsay *et al.* 2001); however, haploinsufficiency of *TBX1* has yet to be proven in humans. Several coding sequence changes have been identified in *TBX1* in patients with DGS (Gong *et al.* 2001), but the pathological significance of these mutations remain unknown at this time (Baldini 2002). In addition, genes that are not haploinsufficient in mice may prove to be haploinsufficient in humans. Although *TBX1* remains the best candidate for causing at least some of the features of DGS, this disorder may also be caused by the deletion of cis-acting regulatory elements or by mutations at other loci (Baldini 2002). These possibilities are currently under investigation.

The only known recurrent constitutional non-Robertsonian translocation in humans involves chromosome 22q11; this translocation occurs between chromosomes 11 and 22. The der(22) syndrome, a congenital anomaly disorder, results from 3:1 meiotic non-disjunction events in carriers of this translocation (Edelmann *et al.*, 2001). Affected individuals carry a derivative chromosome 22, rendering them trisomic for 11q23-qter and 22pter-q11. Clinical features of this disorder include mild

craniofacial anomalies, mental retardation, and congenital heart defects. The derivative 22 chromosome is a duplication of 22q11, but also a duplication of chromosome 11. A pure duplication of the chromosome 22q11 region occurs in patients with cat eye syndrome.

1.3 Cat eye syndrome

Cat eye syndrome (CES) (MIM 115470) is a developmental disorder that derived its name from the strong resemblance of patients' eyes to the shape of the pupil of a cat's eye. The first report of this disorder dates back as early as 1878 when Haab described an association between colobomas (failure of the ocular fissure to close) and anal atresia (failure of the anus to open). However, cat eye syndrome was not described cytogenetically until 1965 when Schachenmann showed an association between a supernumerary, bisatellited marker chromosome in patients and a phenotype that is similar to what is known as the CES phenotype today. In 1981 Schinzel *et al.* concluded the supernumerary marker chromosome associated with this syndrome was an inverted dicentric duplication (Figure 1-1) of a portion of chromosome 22 (inv dup (22) (pter→q11:q11→pter)). McDermid *et al.* confirmed this molecularly in 1986. The CES critical region (CESCR) was originally delineated by the study of ten patients, some of whom had variable sized duplications (Mears *et al.* 1994). Further study of a patient with a unique supernumerary 22 ring chromosome (Mears *et al.*, 1995) narrowed down the CESCR to a 2 Mb region, which spans from the pericentromere of chromosome 22q11.2 to the marker D22S57 (Mears *et al.*, 1995; McDermid *et al.*, 1996).

CES usually results from the presence of a supernumerary bisatellited isodicentric chromosome and hence most patients possess four copies of the CESCR.

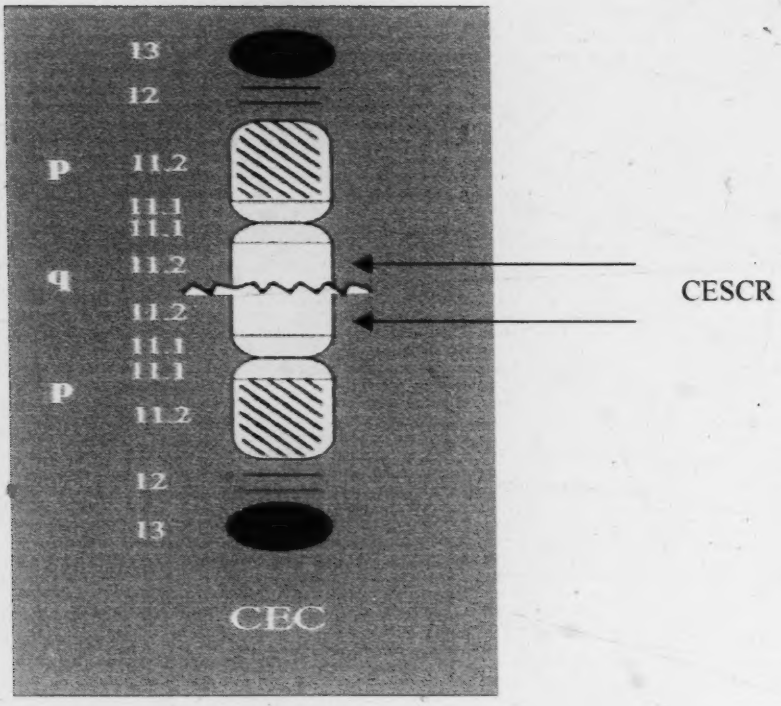


Figure 1-1: The cat eye syndrome marker chromosome (CEC). The typical CEC marker chromosome, as shown above (from Mears 1995), is an inverted isodicentric chromosome that contains two copies of the cat eye syndrome critical region (CESCR). Patients that have two normal copies of chromosome 22 and the marker chromosome possess a total of four copies of the CESCR.

CES has also been reported in patients with an interstitial duplication of the CESCR (Knoll *et al.* 1995; Reiss *et al.* 1985), and has rarely been cited in patients without any visible chromosomal anomalies (reviewed by Franklin and Parslow, 1972). In addition, the size of the duplicated fragment (whether the duplication includes the DiGeorge Syndrome region or not) does not appear to be linked to the severity of the phenotype (McTaggart *et al.* 1998). Mosaicism is evident in some patients and is likely due to the early loss of the marker chromosome during post-zygotic divisions. Again, there does not appear to be any clear correlation between phenotype and the number of cells containing the marker chromosome (Berends *et al.*, 2001). Table 1-1 summarizes the variable phenotype noted in CES patients and the percentage of patients presenting with various clinical features of the syndrome (Berends *et al.* 2001).

The phenotype in CES patients varies from near normal to the presence of severe congenital abnormalities. Physical characteristics include ocular coloboma (failure of the ocular fissure to close), congenital heart defects, ear pits and ear tags, anal atresia (absence of an anal opening), absent or hypoplastic kidneys, and mental retardation. Only 41% of patients have the traditional combination of pre-auricular anomalies, anal anomalies, and iris coloboma, making it difficult to properly diagnose CES. The most common defect of this syndrome is the pre-auricular pit or tag, which is also prevalent in the general population. The two most frequently observed heart defects in CES are total anomalous pulmonary venous return (TAPVR) and tetralogy of fallot (TOF). TOF results from the presence of a hole between the two ventricles of the heart (ventral septal defect), which allows for the mixing of oxygen-rich and oxygen-poor blood. Pulmonary stenosis also results in a partial block in the flow of blood to the lungs. In addition, the muscularity of the right ventricle is significantly

increased, as compared to a normal heart muscle. TAPVR is a heart defect in which the vessels in the heart are not connected properly, resulting in inappropriate blood flow from the lungs to the right atrium, rather than the left atrium. TAPVR accounts for 35% of heart defects in CES patients and TOF accounts for 20% of the cardiac anomalies (Berends *et al.*, 2001). However, the most frequently occurring cardiac defect in the general population is TOF (5-7% of congenital heart anomalies). In a study conducted by Berends *et al.* (2001) only 4% of CES patients presented with TOF, whereas 21% were born with TAPVR. This data suggests that, in contrast to TAPVR (which is very rare in the general population), TOF is not one of the more important features of this syndrome.

1.4 Other Disorders with CES-like Phenotype

CES is a rare genetic disorder. Since the phenotype of CES is variable, and often the clinical features are not observed in combination with one another, other disorders share some of the phenotypic characteristics. This overlap of physical traits may lead to confusion in the patient's initial diagnosis. Several of the phenotypic characteristics of CES are also associated with CHARGE (MIM 214800) syndrome, which is quite rare, as well. CHARGE (coloboma of the eye, heart anomaly, choanal atresia, mental retardation, genital anomalies, ear anomalies) is believed to be a multifactorial trait (Pagon *et al.*, 1981), with only 8% of cases being familial. Mutations in the *PAX2* gene were believed to be the cause of some cases of this disorder. However, in a study by Tellier *et al.* (2000) no mutations in the *PAX2* gene were discovered in any of the 34 CHARGE patients tested. Tellier *et al.* speculate that either mutations in the *PAX2* gene are not a common cause of CHARGE

association or that downstream targets and effectors of *PAX2* (which are presently unidentified) may be candidate genes.

VACTERL (MIM 276950) association is another syndrome that has an overlapping phenotype with CES. VACTERL (vertebral anomalies, anal atresia, cardiac malformations, tracheoesopahgeal fistula, renal anomalies, limb anomalies) often occurs sporadically and is not associated with a recognized teratogen. Mice with a defect in the sonic hedgehog (*SHH*) signalling pathway show a spectrum of developmental defects with similarity to the VACTERL phenotype (Kim *et al.*, 2001).

CES patients possessing only a few phenotypic traits may also be mis-classified with Townes-Brocks Syndrome (MIM 107480). This syndrome was first described in 1972 and is characterized by imperforate anus with hand, foot, and ear anomalies. Deafness is also associated with this syndrome. Townes-Brocks syndrome is caused by mutations in the *SALL1* gene, a putative transcription factor.

Due to the low incidence of CES and the variable penetrance of the phenotype, it is likely that patients may be misdiagnosed with a variety of other syndromes. Although the significant variation noted in the phenotype of CES patients suggests that this disorder is genetically complex and is unlikely to result from the duplication of a single gene, it is possible that only one gene in the CESCR is dosage sensitive and is responsible for the variable CES phenotype. Until all of the genes in the CESCR are functionally characterized, the cause of cat eye syndrome cannot be ascertained. Variations in genetic background may play a role in causing the features of cat eye syndrome. Perhaps a single gene in the cat eye syndrome critical region (CESCR) is responsible for causing all of the features of this disorder, but only certain traits present depending on how the gene product of the CESCR interacts with the patient's genetic background. These are important issues that can only be answered

by identifying more affected individuals and examining the differences between the duplications and by identifying the genes involved in the etiology of cat eye syndrome.

1.5 The Cat Eye Syndrome Critical Region (CESCR)

Until 1995 the cat eye syndrome critical region (CESCR) included greater than 3 Mb of sequence, which is a large area to search within for candidate genes. The CESCR (Figure 1-2) was further delineated by Mears *et al.* (1995) by the study of a unique patient with a ring chromosome. This patient had the classic CES phenotype, but died at 17 days of age; therefore, the developmental potential was unknown. Although both the father and grandfather were phenotypically normal, all three individuals possessed a small supernumerary ring chromosome, the size of which varies in each individual. However, the grandfather and father had only three copies of 22q11.2, whereas the child had four copies. Analysis of the affected individual's ring chromosome redefined the CESCR to the first 2 Mb of chromosome 22q (from the centromere to marker D22S57). Further analysis of patient SK with an interstitial duplication of chromosome 22q11.2 allowed a subsequent subdivision of the CESCR into proximal and distal halves (based on the patient's breakpoint) between markers D22S795 and D22S543. SK has many of the features associated with CES, including mental retardation and facial, heart, kidney, and ear malformations. Anal atresia and ocular coloboma are absent in the phenotype, but may be the result of the incomplete penetrance of this syndrome. Therefore, based on these two patients the features of CES should map to the distal half of the 2Mb region.

Clones from a human genomic library were subcloned into a contiguous array of P1-based and bacterial artificial chromosomes. A set of overlapping clones from the CES region was identified (Johnson *et al.* 1999) and sequenced (Dunham *et al.*

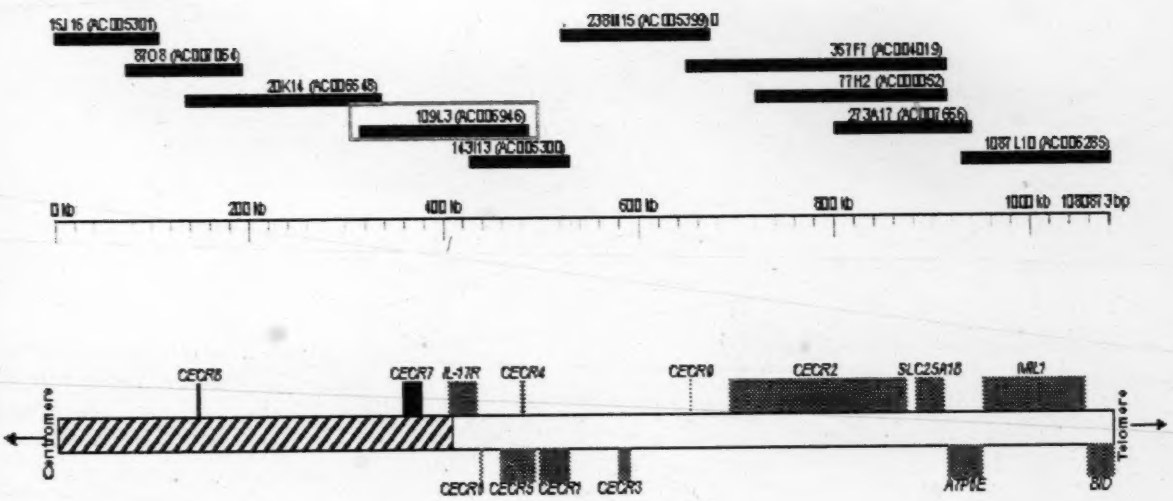


Figure 1-2: A map of the cat eye syndrome critical region (modified from Footz *et al.* 2001). BACs and PACs are shown above the size scale. Genes transcribed centromere to telomere are idenitifed above the chromosome and genes transcribed from telomere to centromere are labeled below the chromosome. PAC109L3, the PAC that contains *CECR7*, is outlined with a red box.

1999); this contig included approximately 1.1 Mb of the distal portion of the CESCR. The contig was partially annotated with known genes (*ATP6E*, *BID*, and *IL-17R*) and various expressed sequence tags (ESTs), and a total of 14 putative genes have been identified to date within this region (Footz *et al.*, 2001). The proximal 400 kb of this 1.1 Mb region (the pericentromeric region) shows a complex pattern of duplicated DNA fragments from various locations within the genome, similar to other pericentromeres (Jackson *et al.*, 1999; Horvath *et al.*, 2000a, 2000b). Two putative genes were identified in this region- *CECR7* (cat eye syndrome critical region gene 7) and *CECR8* (cat eye syndrome critical region gene 8). Both genes produce chromosome 22-specific transcripts, but are comprised of sequence that is highly similar to several different locations in the genome. The discovery of the *CECR7* transcript, derived from several gene sequences from various chromosomes, suggests that the duplication-rich region of the pericentromere may be the birthplace of novel genes formed via the shuffling of exons and genomic DNA. *CECR7* was further characterized and research focused on the evolution of this transcription unit. *CECR8* was also cloned and its expression pattern was determined in humans.

The distal half of the CESCR is gene-rich and contains several candidate genes for involvement in CES. One of the best candidates for involvement in CES is *CECR1*. This gene is hypothesized to function as a growth factor based on its homology to genes in other species. The homologues in several insect species have been shown to regulate extracellular adenosine levels, thereby acting as growth factors. Therefore, the second goal of this thesis was to determine if *CECR1* has ADA activity.

Table 1-1: Summary of the Percentage of Patients Presenting with CES Phenotype

Clinical Feature	**Percentage of CES Patients with Anomaly**
Coloboma	55%
Mental Retardation	32%
Total Kidney Anomalies	31%
Total Anal Anomalies	73%
Total Cardiovascular Anomalies	50%

Table is based on information collected by Berends *et al.* (2001).

Chapter 2: Introduction to *CECR7* and *CECR8*

2.1 The Centromere

When DNA shortens and thickens into chromosomes it becomes visible within the cell. Chromosomes are coiled molecules of double stranded DNA that have separated along the entire length of the molecule, except at the centromere. The centromere, which is comprised of heterochromatic DNA and is important in the process of cell division, acts as the point of attachment for the kinetochore and is responsible for holding homologous pairs of chromosomes together until separation occurs during anaphase. Although the role of the centromere in the segregation of chromosomes in mitotic and meiotic cells is essential and conserved throughout evolution, the structure of centromeres varies among species, and may even vary between chromosomes within a species (Schueler *et al.* 2001).

2.2 Alpha Satellite Repeats

Alpha satellite DNA constitutes 3-5% of all human genomic material (Horvath *et al.*, 2000a) and is highly repetitive sequence. Alpha satellite repeats are comprised of a repetitive 171 base pair unit, the copy number of which varies strikingly between different chromosomes; the arrays formed from these 171 base pair units can vary in length from 200 to 9000 kb (Murphy and Karpen 1998). Some alpha satellites contain a 17 base pair binding site for the centromere-specific DNA binding protein B (CENP-B) and these repeat units are the only repetitive motif to be associated with a postulated function. Several roles have been proposed for alpha satellite repeats (Eichler 1999), including structural or organizational roles at the centromere, a role in chromosome pairing, or involvement in cross-over and recombination. One piece of evidence contributing to the theory that alpha satellite repeats are involved in

centromere function is the ubiquitous presence of these repeat units at the primary constriction of all human and primate chromosomes.

Junctions between satellite DNA and euchromatin are important to study for several reasons (Jackson *et al.*, 1999). Gene transcription can be silenced by repetitive sequences and thus chromatin compartmentalization can be better understood by examining gene regulation/transcription near centromeric repeats. In addition, position effects related to the juxtaposition of constitutive heterochromatin have now been identified in humans and mammals. Therefore, more can be learned about this process, as well (Jackson *et al.* 1999). Satellite DNA is also interesting to study from an evolutionary perspective, as centromeric regions evolve differently than genes on chromosomal arms. Jackson *et al.* (1999) probed human and primate chromosomes with chromosome-specific alpha satellite probes and concluded that the centromeres of these closely related species were strikingly different from one another. The probes they used cross-hybridized to phylogenetically different chromosomes in the various primate species. Results reported by Jackson *et al.* are vastly different than the extensive conserved synteny noted between chromosomal arms in primates and humans.

2.3 The Pericentromeric Region

The pericentromere can be defined as the region of DNA that is distal to the alpha satellite repeat region and extends into the first distinguishable cytogenetic Giemsa stained band on both sides of the centromere (Eichler *et al.*, 1999). Pericentromeres are likely the most difficult region of a chromosome to sequence due to their repetitive nature. Sequencing results for any given pericentromere will likely overlap with sequence from numerous chromosomes (since the pericentromere will

contain duplicons from various genomic locations), which may cause difficulty in assigning clones to one particular chromosome. Until recently pericentromeric regions were presumed to be functionally insignificant. However, putative genes have now been described in the pericentromeric regions of the human genome (Section 2.6). In the past several years evidence has surfaced indicating that pericentromeres are preferential sites for the integration of duplicated fragments of DNA and recent studies suggest the evolution of pericentromeres is surprisingly complex (Eichler *et al.*, 1997, 1999; Jackson *et al.* 1999). Pericentromeres are comprised of duplicons, which can be described as low-copy repeats that show significant paralogy to other genomic locations. At first glance these regions appear to be a junkyard for genic fragments, but they may actually be the birthplace of novel genes.

2.4 Mechanisms of Genome Evolution

Two major molecular mechanisms involved in genome evolution are single-base-pair mutations and gene duplication. New genes with specialized functions can arise after a gene is duplicated and subsequently undergoes mutations. Historically, two different molecular mechanisms have been deemed responsible for gene duplication. The first mechanism (Smith, 1976) involves unequal crossover events that lead to the tandem duplication of an ancestral gene and a family of genes that are very similar to one another. As the species evolves, each family member is susceptible to mutations and members of the gene family slowly diverge and acquire new functions. In contrast, they may lose function and become pseudogenes. The other mechanism by which gene-duplication has been explained is by whole-genome duplication (polyploidy). After the whole genome undergoes duplication,

chromosomal rearrangement occurs and is followed by re-establishment of the disomic state. This model, proposed by Susumu Ohno (1970), has been used to explain the presence of large paralogous regions on non-homologous chromosomes.

However, the genetic consequences of polyploidy in humans are harsh (in most cases is lethal) and thus genome evolution in higher organisms cannot simply be explained by tetraploidization. Since the last tetraploidization event leading to genome duplication in the human lineage is estimated to have occurred greater than 400 million years ago (Lundin, 1993), one would expect that interchromosomal paralogs of recent origin would not be present in the human genome. Since this is not the case, there must be at least one more mechanism by which genome duplication can occur. Sequencing of the human genome has lead to the identification of a third mechanism by which whole genes or large segments of DNA can be duplicated and transposed between non-homologous chromosomes. Genome sequence indicates that large genic segments (that range in length from 5-30 kb) are duplicated within pericentromeric regions of human chromosomes. Eichler *et al.* (1996, 1997) have shown that tetraploidization events cannot account for these genic duplications based on the estimate of the evolutionary age (approximately 33 million years old) of these pericentromeric duplication events.

Eichler *et al.* (1999) identified a CAGGG repeat sequence at the junction of the duplicated sequence (on chromosome 16) that they were studying and proposed that these repeat elements may function in mediating the integration of duplicated fragments of DNA into pericentromeres. Eichler *et al.* (1999) used Fluorescent *In Situ* Hybridization (FISH) to identify the locations of the CAGGG repeat elements and found all signals to be exclusively pericentromeric (using a previously sequenced cosmid from 16p11.2 that contained these CAGGG repeats). CAGGG repeats were

also identified in other primates using FISH and were found to be present at various pericentromeric locations in the hominoid species, but only hybridized to one single pericentromeric location in the Old World Monkeys (macaque). This result suggests that the CAGGG repeat element underwent extensive amplification early on in the evolution of the Anthropoid lineage (Eichler *et al.*, 1999). Furthermore, five known genes (which were duplicated from various locations within the genome) were examined and found to be specifically associated with the CAGGG repeat element, suggesting this CAGGG repeat acts as a signal for the integration of duplicated fragments of DNA in pericentromeres. All of these pieces of data indicate a fairly recent mechanism evolved for the duplication and integration of genic fragments between non-homologous chromosomes.

2.5 Examples of Pericentromeric Duplications

2.5.1 *NF1*

Regnier *et al.* (1997) were one of the first groups to demonstrate the phenomenon of the pericentromere acting as a hotspot for the integration of duplicons while studying the Neurofibromatosis type I gene family. Neurofibromatosis type I is a genetic disorder that is caused by mutations in the *NF1* gene. *NF1* is defined by several unusual characteristics. It is a large gene with an extremely high *de novo* mutation rate and is accompanied in the genome by several other *NF1*-related sequences (Regnier *et al.*, 1997). *NF1*-related sequences are found in mammals, *Drosophila* and bacteria. Some of these related sequences are part of processed pseudogenes, whereas others belong to families of functional genes with analogous roles. Regnier *et al.* (1997) studied the *NF1*-related sequences in humans to better understand the molecular events that led to the creation of this family of related sequences. FISH results indicated that *NF1*-related sequences were present at nine

different loci in the human genome, six of which were pericentromeric and one of which was in the 2q21 region, which is a relic of an ancient centromere (Regnier *et al.* 1997). The extent of the duplications varied between loci as some of the related sequences were much larger than others. Regnier *et al.* concluded that the duplication events were recent, based on comparative sequence analysis with the mouse genome, which did not contain any *NF1* duplications. The majority of the duplications were noted to be in the pericentromeric regions of various human autosomes and thus a pericentromeric specific mechanism of duplication was deemed possible, but not identified by this group.

Regnier *et al.* (1997) also employed PCR to confirm the dating and history of the *NF1* duplications. Primers that were previously used for amplification of the human *NF1* gene were employed to amplify DNA from gibbon and macaque, which were known to have diverged from the human lineage 22 million years ago and 33 million years ago, respectively (Regnier *et al.* 1997). A single PCR product was obtained in macaque, and when sequenced was shown to be highly similar to the functional human gene. Three bands were obtained using the same primers in gibbon. When these bands were sequenced, one was also highly similar to the functional human gene, whereas the other two PCR products were similar to other human *NF1*-related sequences. To rule out the possibility that nucleotide changes in the macaque genome prevented amplification of *NF1*-related sequences, Regnier *et al.* utilized a total of five different primer sets, and each primer set produced only one band, which consistently showed greater similarity to the functional human gene than to the *NF-1* related sequences. In addition, FISH was performed on two different macaque species (*Macaca fuscata* and *Macaca sylvanna*) using the whole human *NF1* cDNA. Results revealed a unique locus on the monkey chromosome homologous to human

chromosome 17, suggesting no *NF1*-related sequences exist in the macaque genome, and that the duplication of this gene began after the divergence of macaque from the Great Ape lineage.

2.5.2 Chromosome 10

Jackson *et al.* (1999) have studied the duplication-rich pericentromeric region of chromosome 10. Analysis of a 9.75Mb YAC-based map across the centromere of human chromosome 10 indicated that duplicated zinc finger (ZNF) gene clusters (which were 150 and 250 kb in size) were linked to centromeric satellites on both chromosomal arms. The duplicated loci were approximately 96% identical at the nucleotide level to one another, suggesting the duplications are quite recent (FISH analyses indicated that the duplications occurred before the divergence of orangutan). FISH analyses of higher primates demonstrated that the blocks of duplicated sequence have undergone many rearrangements over several million years. However, Jackson *et al.* were unable to determine the evolutionary point at which the duplications occurred due to the vast number of rearrangements. They concluded that the pericentromere of chromosome 10 is rich in blocks of duplicated sequences that have undergone many rearrangement events over the last several million years.

2.5.3 16p11.1 Duplication

Eichler *et al.* (1996) identified and analysed an interchromosomal paralogous region between the human pericentromeric region 16p11.1 and chromosome Xq28 (which is not pericentromeric). Analysis of 26.5kb of duplicated DNA and 12.5kb of flanking sequences was performed. The overall sequence identity between the two sequences is 94.6%, which suggests the duplication event occurred fairly recently (7-10 million years ago). At least three genes (the creatine transporter gene, *SLC6A8* and *DXS1357E*) are included in the duplication. FISH analyses of higher primates

indicated that the Xq28 copy is the ancestral copy and that the duplication occurred after the divergence of gorilla from orangutan. Duplications of Xq28 were observed at several different pericentromeres in humans, chimpanzees, and gorillas.

Based on estimates of mutation rates for silent substitutions and for intronic sequences, Eichler *et al.* estimate the duplication between Xq28 and 16p11.1 to have occurred as recently as seven to ten million years ago. However, these calculations are based on mutation rate equations for orthologous sequences in ancestrally related species (for genes in chromosomal arms) based on the time the species was predicted to have diverged from the line leading to humans. The duplication and integration events creating the chromosome 16 copy of the region of Xq28 resulted in a unique chromosomal environment and, therefore, the selection pressure may not be as great as one would predict for an orthologous sequence in a different species. Eichler *et al.* (1996) suggest this duplication may have occurred as recently as four million years ago based on a comparison of the rate of divergence of the beta-globin gene and its paralogous copies in the human genome. The presence of pericentromeric duplications may provide researchers with a novel opportunity to examine the rate of mutation between non-homologous chromosomes and provide further insight into whole-genome evolution. Interestingly, the rate of divergence varies throughout the entire 26.5kb segment. There is a greater degree of sequence conservation in *SLC6A8* exons than in the intronic and intergenic sequence, even though RT-PCR assays specific for chromosome 16 *SLC6A8* transcripts have not identified expression in various tissues tested. However, the normal stop codon is not present in the chromosome 16 copy of *SLC6A8*. Unlike the *SLC6A8* gene, the coding region of *DXS1357E* has diverged fairly rapidly from the paralogous sequence and these base changes have resulted in an altered amino acid sequence and a frame shift mutation in

the pericentromeric copy of this gene. The chromosome 16 paralogous sequence of *DXS1357E* breaks within an intron of the coding sequence, suggesting *DXS1357E* is a non-processed pseudogene. Since only a portion of the coding sequence was duplicated to the pericentromere of chromosome 16, it is likely that the duplication created a non-functional chromosome 16 copy of this gene prior to the occurrence of any sequence divergence. *SLC6A8*, however, may have played a functional role during hominoid evolution, which may explain the apparent tendency to conserve coding sequence.

Further analysis of chromosome 16p11 was performed by Horvath *et al.* (2000a) and revealed that the pericentromere of chromosome 16 has been a hotspot for the integration of duplicated DNA fragments over the last ten million years. Chromosome 16p11 is comprised of genomic fragments from at least two other chromosomal locations. Approximately six million years ago (based on FISH analyses of higher primates) segments from 4q24 and Xq28 were duplicated to the pericentromere of chromosome 16 (Figure 2-1). Horvath *et al.* (2000b) postulate that once this particular duplication occurred the cassette created on chromosome 16 (formed from the genomic segments from chromosomes four and X) was then duplicated to the pericentromeres of chromosomes 2p11, 10p11, 16p11 and 22q11.

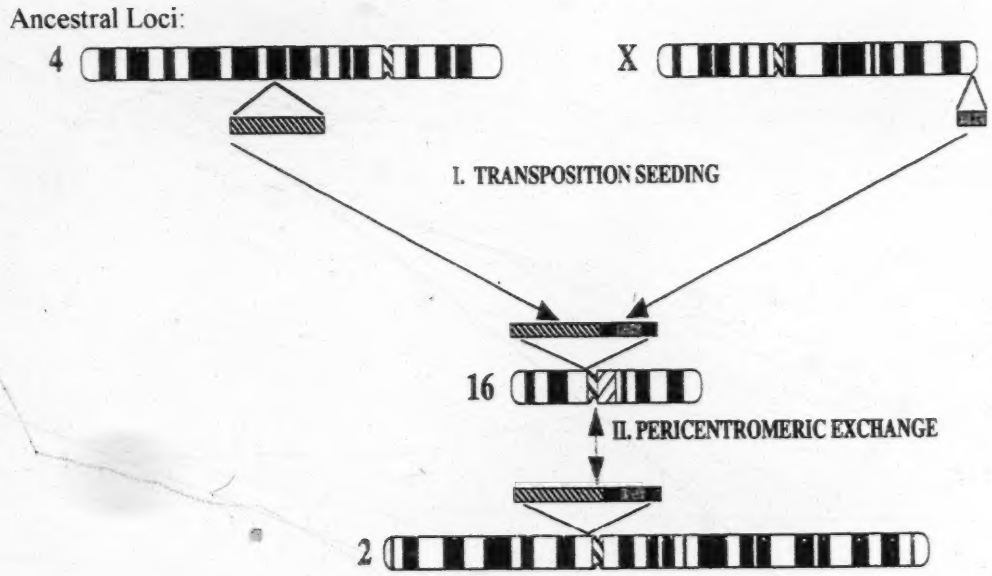


Figure 2-1: Summary of duplication events that occurred between chromosomes 4, X, and 16 (modified from Horvath *et al.* 2000b). Independent duplication (I) of regions from chromosome 4 and chromosome X occurred and seeded the chromosome 16 pericentromere. The larger duplicon was then spread (II) to other pericentromeres.

2.6 Examples of Novel Genes Created via Duplicons

2.6.1 *ASMTL*

Over almost the entire length of the human Y chromosome the frequency of recombination is very low, with the exception of the pseudoautosomal regions (PAR1 and PAR2), in which recombination rates are extremely high. Therefore, the pseudoautosomal regions are duplication-rich. These duplications are believed to be a consequence of unequal crossing over with the equivalent region of the X chromosome. As of 1998, nine genes and one pseudogene had been characterized within the PAR1 (Ried *et al.*, 1998). Only two of these aforementioned genes (*ASMT* and *XE7*) were characterized as "single genes", as all the others have been shown to have close relatives in the genome, or within PAR1. However, Ried *et al.* (1998) isolated a novel gene from the pseudoautosomal region of the Y chromosome called *ASMTL* (N-acetylserotonin *O*-methyltransferase-like). *ASMT* (N-acetylserotonin *O*-methyltransferase) is an enzyme that is involved in the last step of the synthesis of melatonin (a hormone that facilitates sleep) and is expressed solely in the pineal gland, retina, and brain. *ASMTL* shows a high degree of homology to *ASMT* in its' 3' end of the gene. The N-terminus of this novel gene (*ASMTL*), however, is highly similar to the full-length protein of the multicopy associated filament (maf) gene found in *Bacillus subtilus* and to the *Escherichia coli* orfE protein. These bacterial gene products are involved in regulating cell growth and cell division. The full-length *ASMTL* cDNA was transcribed and translated *in vitro* and two proteins of 63 and 70 kDa were observed. The N-terminus (exons 1-6) of *ASMTL* showed similarity to the bacterial genes and the C-terminus (exons 8-13) was highly similar to *ASMT*, as expected. A stretch of 80 amino acids (exon 7) was identified between the two completely diverse N-terminal and C-terminal domains that did not show significant similarity to any known sequence. Interestingly, not all splice junction sites were

conserved in *ASMTL*. In *ASMT* there is an additional intron that interrupts *ASMTL* exon 9, and an additional exon that encodes a portion of a LINE 1 element in *ASMT* that is not present in the putative gene that arose via duplication. In addition, expression studies of *ASMTL* indicate a ubiquitous expression of this gene, suggesting that this novel gene may not function in the melatonin pathway, as does *ASMT*. Therefore, *ASMTL* appears to be a novel gene that has formed in the pseudoautosomal region of the Y chromosome as a result of the fusion of two vastly different full-length genes with different evolutionary origins.

2.6.2 The Drosophila Gene *jingwei*

A similar gene was created in *Drosophila* via the fusion of two unique genes. The *jingwei* (*jgw*) gene exists in only two *Drosophila* species- *Drosophila teissieri* and *Drosophila yakuba*. These two fly species diverged less than 2.5 million years ago. The 3' end of *jingwei* was first identified in 1982 by Langley *et al.* and was subsequently cloned in 1991 by Jeffs and Ashburner. These groups identified the cloned sequence to be a retrosequence that was formed from the alcohol dehydrogenase gene (*Adh*); the retrosequence contained three *Adh* exons, but had lost the intervening introns. Long and Langley (1993) analysed the sequence changes between *jgw* in the two Drosophila species and found most of the changes were synonymous, suggesting the protein sequence was being selectively maintained and was therefore likely functional. RT-PCR was employed to identify *jgw* transcripts and provided further evidence for the functionality of this novel gene. The 5' end of *jgw* is comprised of a group of exons and introns that were recruited by the *Adh* retrosequence into its transcriptional unit to form the chimeric gene structure. After several experiments investigating the 5' end of this chimeric gene, Long *et al.* (1999) concluded that the three 5' exons were from an unrelated pre-existing gene and were

shuffled around the genome and incorporated into a particular genomic location, and thus the *jgw* gene was formed. Since this gene is novel and retains significant features of its early stages, it is possible to study the origin of the *jgw* gene and to better understand gene evolution.

2.6.3 *CT2*- Creatine Transporter 2

These aforementioned genes were formed via the duplication and integration of DNA fragments into particular genomic locations, but did not arise within a pericentromere. Iyer *et al.* (1996) identified a creatine transporter gene at 16p11.2, an example of a pericentromeric duplication of an entire gene followed by evolution of a different expression pattern. This gene, *CT2*, is a duplication of *CT1*, which is located at Xq28. PCR results indicated that there was no size difference in the two cDNAs. *CT1* and *CT2* are 97% identical at the nucleotide level; however, one of these nucleotide changes involved the termination codon. A transversion mutation extends the coding sequence of *CT2* by an additional 50 amino acids. Iyer *et al.* used the additional 50 amino acids at the C-terminus to examine differences in the expression patterns of the two creatine transporter genes. Restriction enzymes that distinguish *CT1* from *CT2* were utilized to digest RT-PCR products obtained from various human tissues. *CT1* was found to be expressed in brain, heart, thymus, intestine, skeletal muscle, testis, and liver. In contrast, testis-specific expression of *CT2* was observed. Iyer *et al.* (1996) postulate that autosomal homologues of X-linked genes exist in order to compensate for the inactivation of genes on the X chromosome in spermatozoa (prior to meiosis). Another examples of this phenomenon involves the phosphoglycerate kinase gene, which maps to Xq13, and its homologue on chromosome 19, which is expressed exclusively in testis.

Since the origin of most genes that we know of today has been obscured by the very evolutionary processes that we are interested in studying, analysis of putative genes such as *CECR7* creates the opportunity to discover the mechanism by which novel genes are created. Whether *CECR7* functions as a protein, an RNA, or is simply a fortuitous transcript that arose in the pericentromere of chromosome 22 remains unknown. However, since the origin of this transcription unit is still traceable, it is important to study in order to help elucidate the mechanism by which new genes arise and the mechanism by which these genes acquire functional properties. The study of *CECR8* may bring to light mechanisms of genome evolution, as well.

2.7 Research Objectives

There were two main projects involved in the completion of this thesis. The first aspect of this thesis involved characterizing the structure and evolution of the *CECR7* gene and characterizing the structure and expression pattern of *CECR8*. This portion of the project involved:

(1) Completion of the cloning of *CECR7*:

Determine the structure of *CECR7* using RT-PCR, 5' and 3' RACE, and the sequence of ESTs.

(2) Identification of paralogous *CECR7*-like sequence in the human genome:

Identify *CECR7*-like sequences within the human genome to determine the origin of the duplicons that combined to form this chimeric transcription unit.

(3) Determination of the time point when *CECR7* was created:

Determine the point in the evolution of the primates at which *CECR7* arose.

(4) *CECR7* expression analysis in humans and primates:

Determine the tissue-specific expression of *CECR7* in humans and examine the expression of *CECR7* in various primate species.

(5) Clone *CECR8*:

Determine the structure of *CECR8* and examine alternative splicing.

(6) *CECR8* Expression Analysis in Humans:

Examine the tissue-specific expression pattern of *CECR8*.

Chapter 3: *CECR7* and *CECR8* Materials and Methods

3.1 Computer Analysis: Identification and Analysis of *CECR7* and *CECR8*

Ali Riazi of the McDermid lab identified *CECR7* by exon trapping *CECR7* exons 2 and 3. This project began with subsequent BLAST (Altschul *et al.* 1997) searches that indicated exons 2 and 3 were 86% and 84% similar to sequence on chromosome 16. Sequence was then repeat masked using the REPEAT MASKER web server (Smit & Green, unpublished data) and entered into the computer program GENSCAN (Burge and Karlin 1997), which predicts gene structure based on sequence and putative splice acceptor and donor sites. Searches were also performed on the human EST database to identify cDNAs that mapped to chromosome 22 in the CESCR pericentromeric region. *CECR8* was identified through a database search of the EST database using the repeat masked sequence of all the PACs in the CESCR centromeric to *CECR7*. Two ESTs were identified centromeric to *CECR7* (hEST 1840041 (accession number AI214826) and 1840206 (accession number AI214704)) and were sequenced. Once *CECR8* was identified, the same *in silico* procedure was followed to elucidate the structure of this putative gene.

3.2 Primer Design

All primers were designed using the computer program GeneTool 1.0. All available *CECR7* primers are listed in Table 3-1.

3.3 PCR: Cloning of Genes and Examination of the Presence of Duplicon Boundaries in Primates

PCR was carried out in a standard buffer (250 mM Tris-HCl pH 9, 500 mM KCl, 15 mM MgCl, 0.2 mg/mL BSA). A 10 mM stock of dNTPs was used and diluted to a final concentration of 0.25 mM. Prior to the PCR reaction, primers were diluted to a concentration of 5 ρmol/μL; the final primer concentration in all PCR reactions was 0.25 ρmol/μL. All reactions included a hotstart and 33 cycles of amplification. The following program was used:

94°C for 1 minute and 30 seconds, 80°C hold, add *Taq*; 33 cycles of: 53°C 10 seconds, 72°C 1 minute, 94°C 10 seconds; 53°C 30 seconds, 72°C 10 minutes, 4°C indefinitely. The annealing temperature of this program was modified for some primer sets and the extension time varied depending on the size of the template to be amplified. Primer sets and the annealing temperatures used are summarized in Table 4-1. The type of *Taq* polymerase used was dependent on the experiment. Commercial *Taq* purchased from BRL was required for PCR across the primate genomic duplication boundaries to minimize background. PCR performed during the cloning of *CECR7* and *CECR8* utilized *Taq* polymerase obtained from the Microbiology Department at the University of Alberta. For these experiments a ratio of 100:1 of *Taq*:*Pfu* was used.

3.4 Identification of Paralogous Locations of *CECR7*-like Sequence: PCR using Monochromosomal Hybrid Panel DNA

PCR was performed as previously described on DNA derived from each of the monochromosomal hybrid cell lines available. The hybrid cell lines were purchased from the Coriell Cell Repository Institute (CCRI). The DNA was diluted to a

concentration of approximately 5 ng/μL prior to PCR, and to a final concentration of 0.1 ng/μL.

3.5 Reverse Transcription PCR (RT-PCR): Expression Analysis of *CECR7* and *CECR8*

RT-PCR was performed using the ThermoScript RT-PCR kit (Gibco BRL) with the following modifications:

a) reaction conditions used for first strand cDNA synthesis were:

   42°C 30 minutes, 50°C 10 minutes, 53°C 10 minutes, 55°C 10 minutes, 57°C 10 minutes, 60°C 10 minutes, 85°C 5 minutes.

b) first strand cDNA products were diluted 1:2 prior to second strand cDNA synthesis

3.6 Rapid Amplification of cDNA Ends (5' and 3' RACE): Cloning of *CECR7* and *CECR8*

5' RACE and 3' RACE (Frohman et al. 1988) were performed according to the SMART™ cDNA Amplification Kit User Manual from CLONTECH, with one minor modification: primary RACE products were diluted 1:49 before proceeding to the secondary reaction. Several RACE libraries were constructed by other members of the McDermid lab using the Marathon™ cDNA Amplification Kit from CLONTECH and these cDNA libraries were utilized in the cloning of *CECR7* and *CECR8*.

3.7 Generation of a Testis cDNA Library

The testis cDNA library was generated using the Marathon™ cDNA Amplification Kit from CLONTECH Laboratories Incorporated. 1 μg of polyA+ testis RNA was used. There were no deviations from the protocol.

3.8 Primer Extension: Identification of the Transcription Start Site of *CECR7*

The following protocol was adapted from Current Protocols in Molecular Biology (1991). A 30 nucleotide reverse primer (*CECR7*-primer extension) was used for primer extension. 100 ng of the primer was end-labelled in a reaction containing 10 mCi/mL of [γ-^{32}P] ATP, 4 units of T4 polynucleotide kinase, and 1X polynucleotide kinase buffer at 37°C for 30 minutes. Heating the reaction to 65°C for five minutes subsequently inhibited the kinase activity. The labelled oligonucleotide was precipitated (a total of three times) with 100% ethanol and 4 M ammonium acetate on dry ice. Cerenkov counting determined the extent to which the primer was radiolabelled. Successful labelling resulted in $3x10^7$ cpm of probe. Approximately 26 µg of kidney RNA was then used for the hybridization. The RNA was incubated with $5x10^4$ counts of labelled probe, ethanol precipitated and resuspended in 1x aqueous hybridization solution (1 M NaCl, 0.17 M HEPES pH 7.5 and 0.33 mM EDTA pH 8.0). Hybridization occurred overnight at 30°C. The following day the hybridization reaction was ethanol precipitated and utilized in the reverse-transcription (RT) reaction. The RT reaction was performed in a total reaction volume of 25 µL and required 0.5 µM dNTPs, 1X RT buffer (50 mM Tris HCl pH 8.0, 5 mM $MgCl_2$, 5 mM DTT, 50 mM KCl and 0.05 mg/mL BSA), 1.25 µL of Rnase OUT (BRL), and DEPC-treated water. 40 units of AMV reverse transcriptase were added to the reaction, which was carried out at 42°C for 90 minutes. The reaction was then stopped by adding 0.5 µM EDTA. Remaining RNA was degraded with 1 µg of Rnase H at 37°C for 30 minutes. To isolate the product, a phenol-chloroform extraction, followed by ethanol precipitation, was performed. Products were subsequently run on a polyacrylamide gel.

3.9 Northern Blot Transfer: Expression Analysis of *CECR7* and *CECR8*

Northern blots were prepared using 5 μg of total RNA as described in Current Protocols in Molecular Biology (1991). The protocol was amended as follows:

a) ethidium bromide was added to the samples prior to loading the gel rather than staining with ethidium bromide afterwards

b) the nylon membrane was exposed to UV light for 30 seconds (to bind the RNA to the membrane) after transferring overnight (in lieu of baking for two hours)

3.10 Northern Hybridization

All probes used in Northern hybridizations were obtained from cDNAs. Probes were labelled using the Ambion Strip-EZ™ DNA Kit with the following modifications:

a) the total reaction volume was reduced to 20 μL

b) half the amount of water, 10X Decamer solution, 10X CTP, and 5X buffer were used

c) half the amount of radiolabelled [α-^{32}P] ATP was used

d) hybridization to Northern blots containing primate RNA was performed at a reduced stringency in order to identify diverged paralogous RNA; Northern blot hybridization occurred at 37°C

e) following hybridization, low and high stringency washes were performed for a variable length of time, usually between five and twenty minutes, depending on the level of background radioactivity

3.11 DNA Purification from Agarose Gels

The GENECLEAN II® Kit (BIO101) was employed to isolate and purify DNA from agarose gels. All agarose gels were made with 1X TAE buffer; therefore, the TBE modifier solution from the kit was not required. The following modifications were made to the protocol:

a) all agarose bands, independent of their mass, were dissolved in 3 M NaI at 50°C

b) 7.5 μL of glass milk was added to all samples purified

c) only one elution was performed to recover DNA bound to the glass milk

3.12 Plasmid DNA Preparation and Purification

All PCR products were subcloned into pGEM®-T Easy or pBluescript II SK+™ and transformed into XL1-blue competent cells by standard procedures (Sambrook and Russell 2001). Single colonies were then grown up overnight in standard LB media at 37°C, in the presence of 0.16 μg/μL ampicillin. 3 mL of bacteria were then isolated by centrifugation and plasmid DNA was extracted from these pellets using the protocol described in the QIAprep® Miniprep Handbook by QIAGEN.

3.13 Sequencing

Sequencing was performed on the ABI prism 377 sequencing machine. Samples were prepared for sequencing as outlined by Applied Biosystems Incorporated for the DYEnamic™ ET Terminator Cycles Sequencing Premix Kit. The protocol was modified as follows:

a) half-reactions were used; therefore, 4 μL of pre-mix, 0.75 μL of pellet paint, 2.25 μL of water and 4 ρmol of primer were required

Approximately 500 ng of plasmid DNA were used in the sequencing reactions.

3.14 Tissue Culture: Growth of Primate Lymphoblast and Fibroblast Cells

All cell lines, both fibroblast and lymphoblast, were cultured in T25 flasks at 37 °C. When fibroblast cells reached confluency, they were washed with 3 mL of serum-free media to remove protein and debris and were then trypsinized. Once the cells lifted from the plate, 10 mL of sterile media (containing the appropriate concentration of FBS) was added and the cells were split 1:4. When lymphoblast cells had grown to a substantial density, excess media was removed by aspiration and the cells were washed and split 1:5. Three cell lines were obtained from the European Collection of Cell Cultures (ECACC). The chimpanzee (*Pan paniscus*) lymphoblast cell line [EB176(JC)] and the gorilla (*Gorilla gorilla*) lymphoblast cell line [EB(JC)] were cultured in RPMI media, 15% fetal bovine serum (FBS), and 1% L-glutamine. These cells were split 1:5 every second day. A kidney fibroblast cell line (PMK) from Rhesus monkey (*Macaca mulatta*) was obtained from ECACC, as well. These cells were grown in EMEM, 5% FBS, and 1% L-glutamine. Two fibroblast cell lines were also obtained from Tamim Shaihk (Children's Hospital of Philadelphia). Both CH99-112F (from the lowland gorilla, *Gorilla gorilla*) and CH99-113F (from macaque, *Macaca mulatta*) were cultured in MEM, 10% FBS and 1% L-glutamine. The orangutan cell line (GM06213) was obtained from the Coriell Cell Repository Institute (CCRI) and was grown in MEM, 20% FBS and 1% L-glutamine.

3.15 Preparation of Freezer Stocks of Cell Lines

The media was aspirated from the cells when they reached confluency; (adherent cells were then trypsinized), cells were washed with media and centrifuged for 10 minutes at 1,000 RPM. Fetal bovine serum (10%) and 10% dimethyl sulfoxide (DMSO) was

added to the appropriate medium. The media was then filtered with a 0.22 μM Nalgene filter. After aspirating the cell pellets, they were resuspended in 1 mL of the filtered media, frozen at -70°C, and stored permanently in liquid nitrogen. Prior to indefinite storage, one stock of each cell line was thawed and grown to ensure viability.

3.16 Isolation of DNA from Cell Lines

DNA was isolated from cells that had been centrifuged down and frozen at -70°C until DNA could be harvested. Cells were resuspended in approximately 5 mL of media (any type) with a sterile pipette and 10 mL of DTAB was then added to the resuspended cells. The cells were incubated at 68°C for 25 minutes and transferred to a tube containing 15 mL of chloroform. The tube was inverted several times and centrifuged at 10,000 RPM for 15 minutes in the HB-4 rotor in the RC5B Sorvall centrifuge. The upper aqueous layer was removed to a new tube containing chloroform and the extraction was repeated three times. Once the top layer was isolated, 15 mL of water and 1.7 mL of CTAB were added, the tube was inverted two times, and then centrifuged at 10,000 RPM. The pellet was resuspended in 6 mM NaCl and left to shake overnight. The following day the DNA was ethanol precipitated and vacuum dried. DNA was resuspended in Tris EDTA pH 8.0.

3.17 Isolation of RNA from Cell Lines

Cell pellets were resuspended in 1 mL of Trizol using a 20-gauge 1-inch needle. After incubation at room temperature for 5 minutes, cells were centrifuged for 10 minutes at 7,000 RPM in the HB-4 rotor in the Sorvall. Chloroform (200 μL) was

added to the cells, they were shaken vigorously for 30 seconds, and then incubated at room temperature for 3 minutes. After centrifugation at 7,000 RPM for 15 minutes, the upper aqueous layer was removed and an equal volume of cold isopropanol was added. The cells were incubated again at room temperature for 10 minutes and then centrifuged for 15 minutes at 7,000 RPM. After the supernatant was decanted, and the pellet dried, 300 μL of RNA isolation buffer (300 μL of solution B and 2.1 μL β-mercaptoethanol) and 600 μL of cold isopropanol were added and the cells were left at room temperature for 10 minutes. Following room temperature incubation, the samples were centrifuged at 7,000 RPM for 15 minutes, the supernatant was decanted and the pellet was dried. The pellet was washed two times with 75% ice cold ethanol, dried and then resuspended in DEPC-treated water.

Table 3-1: List of Available CECR7 Primers

Primer Name	Primer Sequence (5'→3')	Tm (°C)
SAHL 22 F1	ACGTCTGCGACACCGGGGACA	71.4
SAHL 22F2	GAGCAACTGTGGATGAAGAGG	66.4
SAHL RT-R1	CCCCCACCTCAACTCCATTCTGCGTC	73.2
EXON A-2	TCAGCATCCGCGAGGGGGTCC	73.0
INTRON 3-F3	ACTCCTGCCCCATGTCTCCTA	68.1
SAHL EXON 1 R4	CCTCAACTCCATTCTGCGTCA	66.5
SAHL EXON 1 R5	GACCCAGACCCACTCCTTCCC	71.4
INTRON 1-R2	CACAAGGGAAGTTAGAAATTATI	60.3
EXON A-2 NESTED	GACCCAGACCCACTCCTTCCC	71.5
EXON F-F1	TTGCAGCCTGTTCGAAGACTCA	66.5
EXON F-R1	TAGGGTATACTGAGAAGGAAGA	63.4
11-F1	GCATTGATTGTCCTGACCCCAC	68.5
22-R1	CCCCAGGAACCAGATGTAGCCC	71.1

Chapter 4: Results

4.1 Cloning of *CECR7*

CECR7 was identified by Ali Riazi by exon trapping *CECR7* exons 2 and 3 (Figure 4-1) from a cosmid pool containing cosmids c1A3, c1H9, and c88G3. These two exons map to PAC109L3 (Figure 1-2) and respectively show 86% and 93% identity to sequence on chromosome 16 (AC003034) (Riazi 1998). Subsequent database searches with the sequence of these two exons identified two ESTs (hEST 23249 (accession number AA320773) and hEST 1367959 (accession number AA810282)) that were 98%-100% identical to chromosome 22, which were fully sequenced by Tim Footz and Lindsay Bridgland to further elucidate the structure of the gene. Exons 4 and 4'' were identified from the sequence of these two ESTs. Further cDNAs were obtained from a CaCo cDNA library from which the third alternative 3' exon (4') was identified. The GENSCAN computer program was then utilized to predict the gene structure upstream of *CECR7* exon 2, and two exons were predicted.

To further extend the gene sequence, 5' RACE was performed. 5' RACE products were obtained using two different primer sets (Table 4-1). The first strand of the cDNA was synthesized using primers in exon 2 (exon A-1 or SAHL 22-A) and the second strand synthesis was performed with a nested primer in exon 1 (exon A-1 nested or SAHL 22-A nested). Nested primers were utilized to increase the chance of obtaining *CECR7*-specific PCR products, rather than expressed copies of exon 1 or exon 2 resulting from the extensive duplication of *CECR7* exons elsewhere in the genome (see Section 4.3 for the rational). 5' RACE only extended the sequence by

Table 4-1: PCR Primers Used in the Cloning and Characterization of *CECR7* and *CECR8*

Gene Name	Protocol	Primer Name	Primer Sequence (5'→ 3')	Tm (°C)
CECR7	5' RACE	exon A-1	GGCCTGCCACCTCCTCCACC	73.25
		exon A-1 nested	AAACGGAAGCCACCACATCT	64.75
		SAHL 22-A	CGTCCCCACATTCCACTTCG	68.15
		SAHL 22-A nested	GACTCGCTCTGACCACGCCG	71.55
	3'RACE	SAHL RT-F1	CCCGTGGGTGTCACAGAGACCGATAC	69.3
		SAHL 3-1 nested	AGGAGTGGGTCTGGGTCCAGTCCGCA	74.5
	RT-PCR	intron 1-F	AGGACGAATGTAGACGGTGAG	66.5
		intron 1-R	GCTTCCCGGGGCCTGCATG	74.3
	PRIMER EXTENSION	SAHL exon A-primer extension	CTGCCCCCACCTCAACTCCATTCTGCGTCA	74.7
	INTRON 1 DUPLICON BOUNDARY	Intron 1 F3	ATTTTTGTTACCAGAGCAGTT	60
		Intron 1 R3	CTCCCATTTCAGCTCAACCAG	66.5
		Intron 1 F2	CACGCCTGGCTAATTTTTGTATI	63.6
	INTRON 3 DUPLICON BOUNDARY	Intron 3 F2	CGTCACCTTGTTTTTCCACTC	64.9
		Intron 3 R3	TTGCATGTGCTTGTAATATCG	61.7
		Intron 3 F1	TGCACAAATGTTTAAGACCAA	60.0
		Intron 3 R1	CCCTCCTCTGTGTTATTTCAI	63.2
	PRIMATE RT-PCR	Intron 1-F	AGGACGAATGTAGACGGTGAG	66.5
		exon 2-R1	CCCCAGGAACCAGATGTAGCCC	71.1
		SAHL 11-R1	TTTACATGCTGCTGCAGCTCC	66.5
		SAHL 11-R2	CTGGTCCTTGTCATGGGACAG	68.1

Table 4-1 (continued): PCR Primers Used in the Cloning and Characterization of *CECR7* and *CECR8*

CECR 8	RT-PCR	*CECR8* 3F	CTGCTTTTCTAGCCAACTTTC	63.2
		CECR8 2R	GTTGTAACTAGGGTGATATTT	60.0

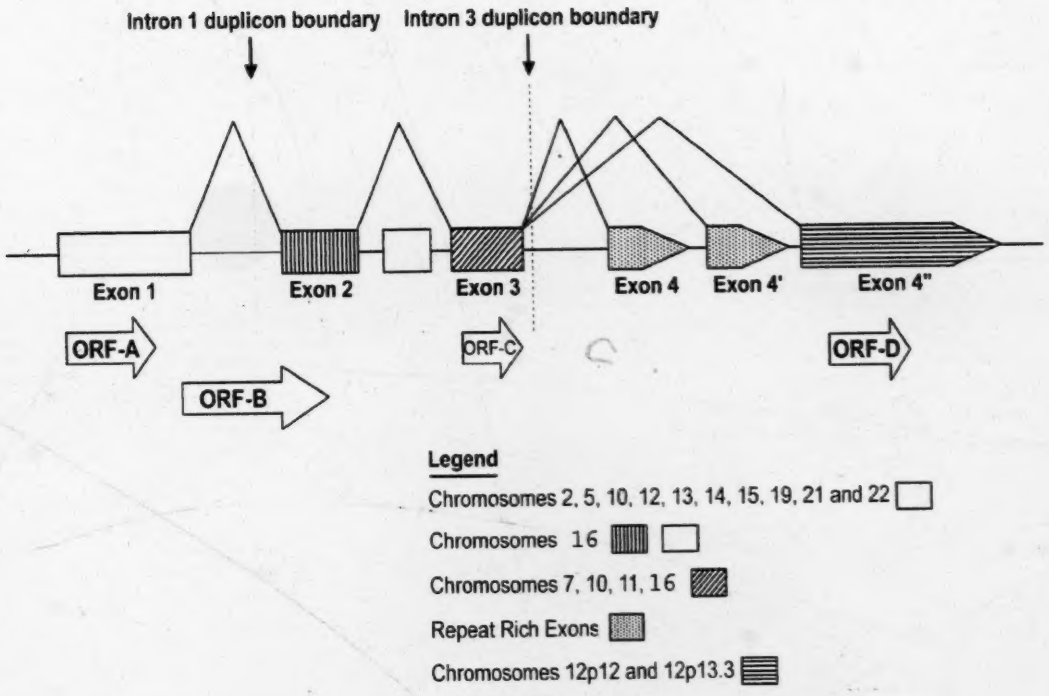


Figure 4-1: Genomic structure of *CECR7*. Regions of similarity, as determined by BLAST analysis and monochromosomal hybrid panel PCR, to other chromosomes are depicted by the various patterns listed below the gene. Duplication boundaries within introns 1 and 3 are shown, as are the four putative open reading frames. Reading frame for ORF-A and ORF-B is 0 and the reading frame for ORF-C and ORF-D is +1.

approximately 25 nucleotides. Therefore, other methods of cloning were required to elucidate the structure of the gene.

Based on gene prediction programs, *CECR7* exon 1 was believed to be two smaller exons; therefore, 3' RACE was performed using primers (SAHL RT-F1 and SAHL 3-1 nested) (Table 4-1) in the predicted upstream exon. RACE products were sequenced and indicated that the two predicted exons were in fact one large exon (the RACE product included a portion of the GENSCAN-predicted intron 1).

Reverse transcription-PCR (RT-PCR) was then performed to confirm the presence of the single 5' exon and to rule out the possibility that genomic DNA contamination was the template from which the 3'RACE products were derived. The cDNA was generated using the oligodT_{20} primer, and primers within predicted intron 1 (intron 1-F and intron 1-R) (Table 4-1) were used for the secondary PCR. A negative control for each reaction was performed. These controls contained RNA, but no reverse transcriptase; therefore, products observed for the negative controls following PCR would indicate that the RNA was contaminated with genomic DNA. However, as demonstrated in Figure 4-2, these negative controls indicated the RNA was not contaminated with genomic DNA. PCR products were sequence verified. Results demonstrated that the first two GENSCAN predicted exons were in fact one exon, as shown in Figure 4-2. RT-PCR was subsequently performed between *CECR7* exons 1 and 2 to confirm these results (Figure 4-14 and Section 4.4).

The known sequence of *CECR7* putative exon 1 is 775 base pairs. No other exons are predicted upstream of this exon by GENSCAN. To identify the transcription start site, primer extension was attempted. Kidney RNA was used in the reverse transcription reaction and the reverse primer (SAHL exon A- primer extension) (Table 4-1) was radiolabelled. Products were run on a manual sequencing

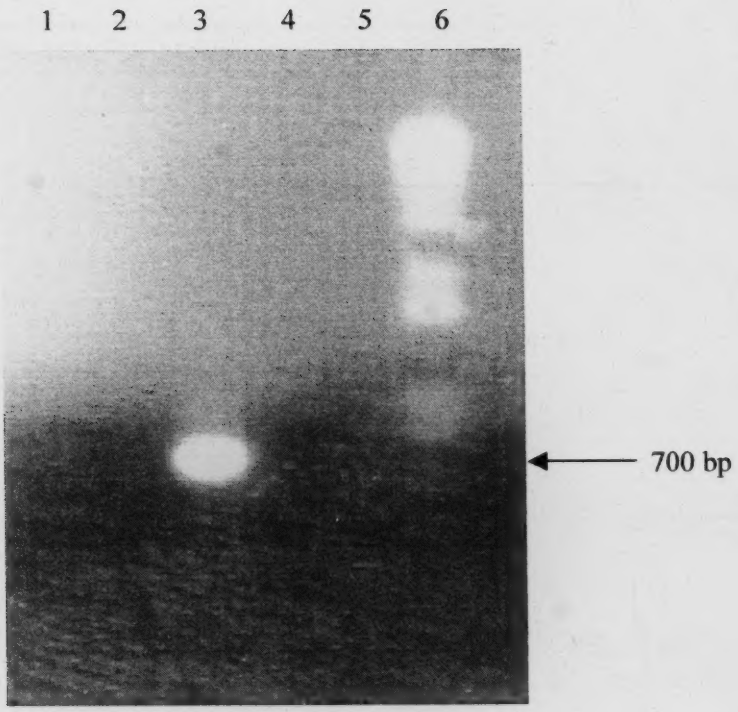


Figure 4-2: Exon 1 RT-PCR results. Results demonstrate that predicted exons 1 and 2 are a single exon. Prior to the reverse transcription reaction, RNA was treated with DNaseI to prevent amplification of products from genomic DNA. RT-PCR results indicate *CECR7* is not expressed in lung, muscle, or heart, but is expressed in kidney. The PCR was performed with the primers SAHL RT-F1 and SAHL 22-A nested. Lane 1: Negative control; Lane 2: Lung; Lane 3: Kidney; Lane 4: Heart; Lane 5: Muscle; Lane 6: 1kb+ ladder

gel and compared to a known sequence. However, primer extension was not useful in determining the transcription start site of *CECR7*. Radiolabelled *CECR7* reverse transcription products were not apparent, although a positive control indicated the protocol was effective. Absence of *CECR7* primer extension products likely resulted from technical difficulties.

Although it was not possible to identify further upstream gene sequence by primer extension or through gene prediction programs, a large open reading frame is not evident in the *CECR7* sequence. *CECR7* possesses four short open reading frames (ORF) (Figure 4-1). ORFs A through D are 76, 103, 61 and 77 amino acids, respectively. Two of the ORFs show similarity to known proteins. ORF-B is 30% similar to the endothelin B receptor protein (AAF81902) in *Canis familiaris*. ORF-C is 79% similar to the *Bos Taurus* lipase activating enzyme (AB04289), 79% similar to a xenobiotic medium-chain fatty acid coenzyme A ligase (AF126145) and is 69% similar to the *Homo sapiens* butyryl coenzyme A synthetase I (BUCS I) gene. Whether any of these open reading frames are translated or whether they are functional remains unknown.

4.2 Expression Analysis of *CECR7*

CECR7 is expressed in limited tissues. The human ESTs sequenced to confirm the structure of this putative gene were derived from tonsil and adipose tissue, demonstrating *CECR7* expression in these tissues. A further expression profile of *CECR7* was elucidated using RT-PCR. Reverse primers in exons 2, 3, and 4'', as well as the oligodT20 primer, were all employed in an attempt to determine the spatial expression of *CECR7*. RT-PCR products were not obtained from RNA derived from several tissues, including lung, liver, muscle, testis, and lymphoblasts (Figure 4-2,

Figure 4-14). Perhaps *CECR7* is expressed in these tissues at such a low level that expression was not identified by RT-PCR, or perhaps it is not expressed. *CECR7* RT-PCR products were repeatedly obtained from human kidney RNA and fibroblast RNA.

Northern blot analysis was unsuccessful in further determining or confirming the expression pattern of *CECR7*. Various probes were utilized to determine expression of this putative gene in several adult and fetal tissues, including adult kidney and fibroblasts (tissues for which expression had been confirmed by RT-PCR). Probes were generated to exon 1, exon 1-2, exon 1-3, and exon 4''. However, *CECR7* was not apparent on any of the Northern blots tested. Since a *CECR7* transcript was not identified by Northern blot analysis, it is not possible to conclude whether the transcription start site is contained within the known *CECR7* sequence (as discussed in Section 4.1). Control probes demonstrated even loading of RNA, indicating that the lack of Northern results was likely due to a low level of *CECR7* expression.

4.3 Identification of *CECR7*-like Sequences Throughout the Human Genome

BLAST analysis of repeat-masked sequence and sequencing of PCR products derived from monochromosomal hybrid cell lines (NIGMS panel #2) indicated that *CECR7* exons 1, 2+3, and 4 were independently duplicated (on separate duplicons) and are highly similar to sequence on many different human chromosomes (Figure 4-1). To ensure that the *CECR7* set of duplicons were not present together as a unit on other chromosomes, PCR analysis of monochromosomal hybrid panel DNAs was performed using primers flanking the genomic duplicon boundaries (Figure 4-1). The boundary between the duplicon containing *CECR7* exon 1 and the duplicon containing exons 2 and 3 lies within intron 1. The boundary between the exon 2 and 3

duplicon and the exon 4 duplicon lies within intron 3 of *CECR7*. Primers were designed to flank the two human duplicon boundaries. Amplification of these boundary regions only occurred in the chromosome 22 monochromosomal hybrid and the control chromosome 22 PAC clone (109L3, AC006946) containing *CECR7* (Figure 4-3). Thus, these results are consistent with the *CECR7* transcript being unique to the chromosome 22 pericentromeric region, even though the individual duplicons occur in many places throughout the genome.

Although the hybrid panel was employed to confirm *CECR7* chromosome 22 specificity, it was also useful in verifying and further identifying *CECR7* exon 1-like sequence throughout the human genome. Primers flanking *CECR7* exon 1 were used to perform PCR on the monochromosomal hybrid panel. Despite the expected variation in sequence between paralogous copies of *CECR7* exon 1, PCR products were obtained for several different chromosomes (Figure 4-4). Although sequence reads of the PCR products were too poor to determine the exact degree of similarity between *CECR7* exon 1 and paralogous sequences, they did confirm that the PCR products were in fact similar to *CECR7* exon 1. In addition, not all known locations of *CECR7* exon 1-like sequence were identified by PCR (See Discussion Section 5.1). Results from the hybrid panel identified paralogous sequences on chromosomes 5 and 19 (Figure 4-4), which were not known based on BLAST analysis, suggesting that pericentromeric regions may be underrepresented in the NCBI nr and htgs databases. A PCR product of the predicted size (750 bp) was also observed for chromosome 3; however, this product could not be verified by direct sequencing of PCR products. Whether sequence on chromosome 3 is similar to *CECR7* exon 1 still remains unclear. BLAST analysis also indicates exon 1 shows similarity to sequence on chromosomes

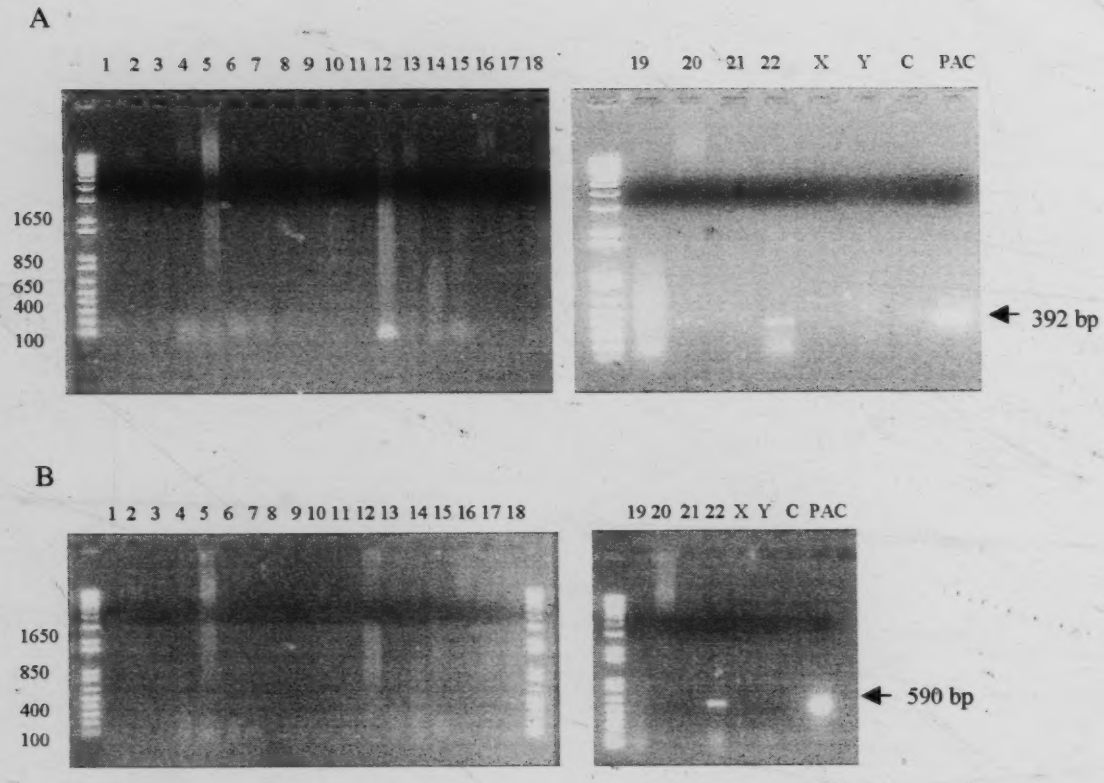


Figure 4-3: Monochromosomal hybrid panel PCR results spanning *CECR7* intron 1 and intron 3 duplicon boundaries. Results indicate that *CECR7* duplicons are not present as a unit on any chromosome other than chromosome 22. PAC109L3 was used as a positive control, as was chromosome 22 DNA. Panel A demonstrates results obtained across the intron 1 duplicon boundary and panel B shows results obtained across the intron 3 duplicon boundary. The numbers above the individual lanes identify the chromosome from which the DNA was derived.

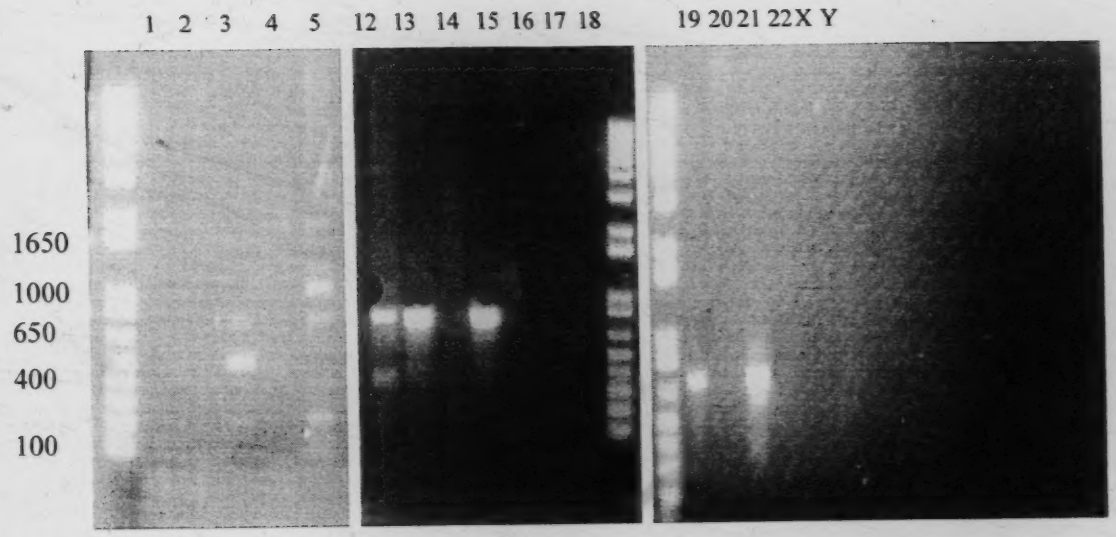


Figure 4-4: NIGMS panel #2. PCR was performed on the monochromosomal hybrid panel DNA to identify which chromosomes possess *CECR7* exon 1-like sequence. PCR results identified exon 1-like sequences on chromosome 5, 12, 13, 15, 19, and 21. The smaller band (450 bp) for chromosome 3 was sequenced and did not show similarity to chromosome 22. Chromosomes 6-11 were also tested, but PCR products were not obtained (data not shown). Selected ladder sizes are labeled on the left side of the diagram.

2, 10, 12, 13, 14, 15, 21, and elsewhere on chromosome 22. Monochromosomal hybrid panel PCR was not useful in identifying paralogues of *CECR7* exons 2, 3 (as will be discussed), or 4. PCR was performed using various primers within exon 4'', but consistent results for known copies of exon 4'' (ie. chromosome 12 and 22) could not be obtained. In addition, exon 4''-like sequence was not identified on any other chromosomes.

The origin of the exon 1 duplicon is unknown, despite several attempts to identify its original location by BLAST analysis. The exon 1-like sequences on chromosomes 2, 14, and another region of 22 are the least similar to *CECR7*, suggesting a more evolutionarily distant duplication, but BLAST search analysis did not identify which, if any, of these sequences are non-pericentromeric. Several ESTs were identified through BLAST analysis that are 85-90% similar to *CECR7* exon 1, suggesting an expressed copy of exon 1 exists in the genome. Two of these ESTs, hEST 4668975 (accession number 18593585) and hEST aX10h05.x1 (accession number BG941273), which are only similar to the last 100 nucleotides of *CECR7* exon 1, correspond to the "hypothetical gene supported by AK056135 mRNA" (accession number XM_086797). Human EST 4668975 was obtained from prostate tumour cells and hEST aX10h05.x1 was derived from blood cells. BLAST analysis of AK056135 identified two chromosomes to which the sequence of this mRNA is 100% identical. AK056135 maps to chromosome 14 (clone R-146E13, accession number AL589182.3) and to a location on chromosome 22 that is telomeric to the cat eye syndrome critical region (clone RP11-623D21, accession number AC080181). This hypothetical gene is also 86-91% similar to sequence on chromosome 15 (clone RP11-536E9, accession number AC022372), and is present in its entirety, with the intervening genomic sequence between the predicted exons. The chromosome 14 or

the chromosome 22 copy of this hypothetical gene is possibly the origin of *CECR7* exon 1. Although only the first 100 nucleotides of AK056135 (predicted exon 1) are similar to the 3' end of *CECR7* exon 1 (87% similarity), the 600 nucleotides of genomic DNA directly upstream of the hypothetical gene exon 1 are similar to the remaining 5' end of *CECR7* exon 1 (Figure 4-5). Perhaps a genomic duplication of this region of DNA, which includes the "hypothetical gene supported by AK056135 mRNA", occurred on chromosome 14 or chromosome 22, spread to a pericentromere and further, smaller, duplications then spread throughout the genome. Interestingly, the putative exons of AK056135 show similarity to other locations on the same chromosome 14 and chromosome 22 clones with a lower degree of similarity (87-91%), suggesting portions of this region were also duplicated to non-pericentromeric loci on chromosomes 14 and 22. Perhaps this region of the genome is highly susceptible to duplication.

The sequence flanking either side of exon 1 shows a high degree of similarity (over regions greater than 1 kb in size) to chromosomes 1, 2, 4, 6, 7, 10, 12, 13, 14, 15, 21, and elsewhere on 22. Figure 4-6 illustrates the complex nature of the DNA flanking exon 1 and the other *CECR7* exons.

Based on BLAST analysis, exon 2 is similar to sequence on chromosome 16. Exon 3-like sequence is present on chromosomes 7, 10, 11, and many locations on chromosome 16. PCR was performed on the monochromosomal hybrid panel to determine further paralogous locations of these two exons; however, due to their small size, PCR products could not be successfully differentiated from primer dimers. PCR between exons 2 and 3 was unsuccessful, as the region is approximately 2500 base pairs in size and was too large to amplify. In addition, PCR between these two exons would only identify chromosomes on which exons 2 and 3 integrated as a

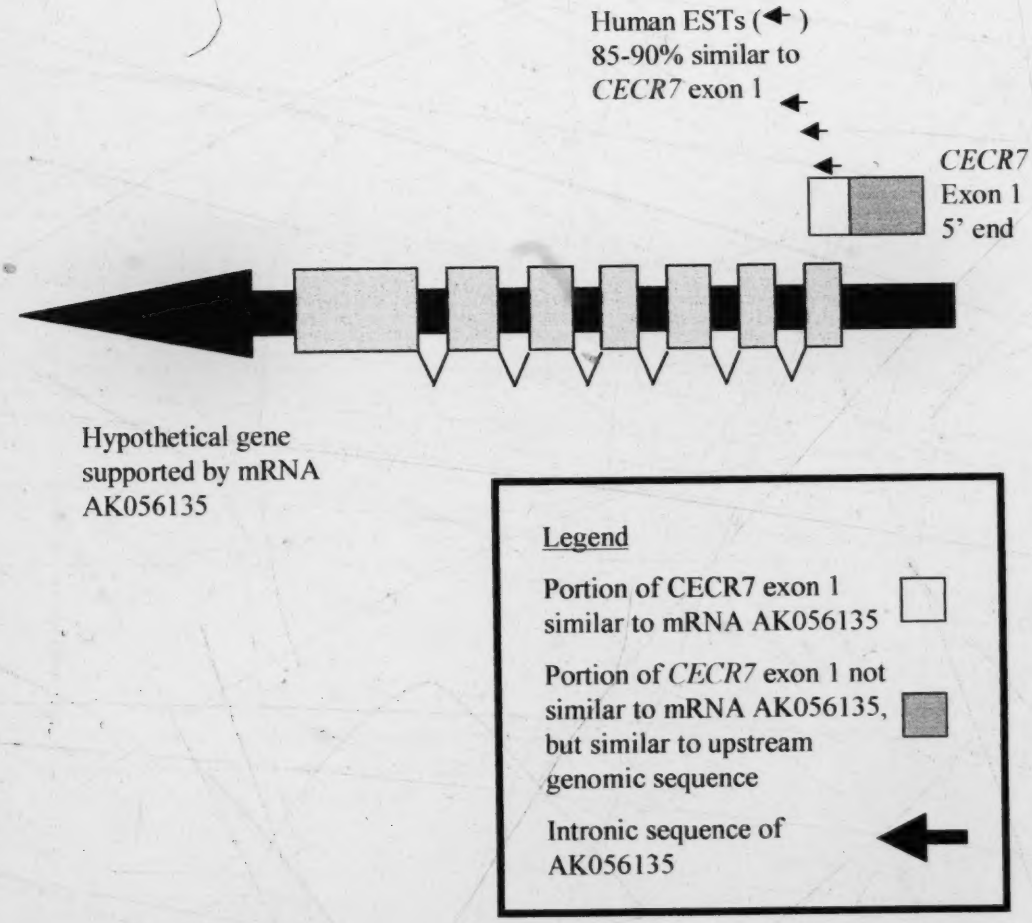
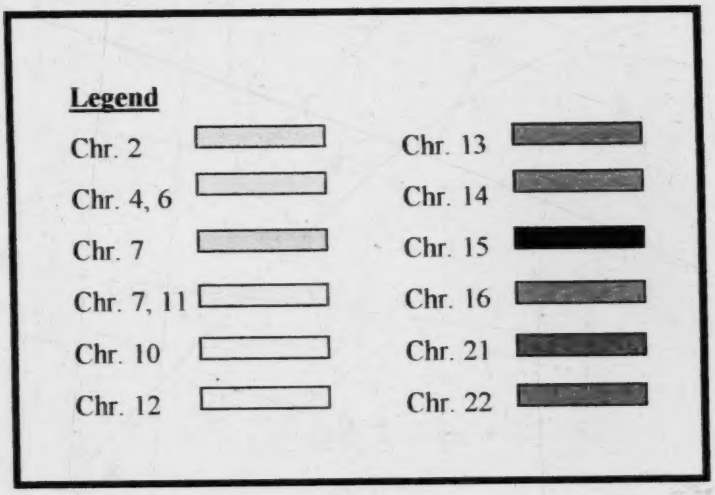


Figure 4-5: Diagram illustrating the similarity of *CECR7* exon 1 to a hypothetical gene represented by several human ESTs that are 85-90% similar to the 3' end of exon 1. Coloured boxes represent the putative exons of the gene supported by mRNA AK056135. The 3' end of *CECR7* exon 1 is similar to the first exon of the hypothetical gene. The remaining 600 nucleotides of the 5' end of *CECR7* exon 1 are similar to the genomic sequence directly upstream of exon 1 of the hypothetical gene. The genomic region containing this hypothetical gene is 100% identical to chromosomes 14 and 22.



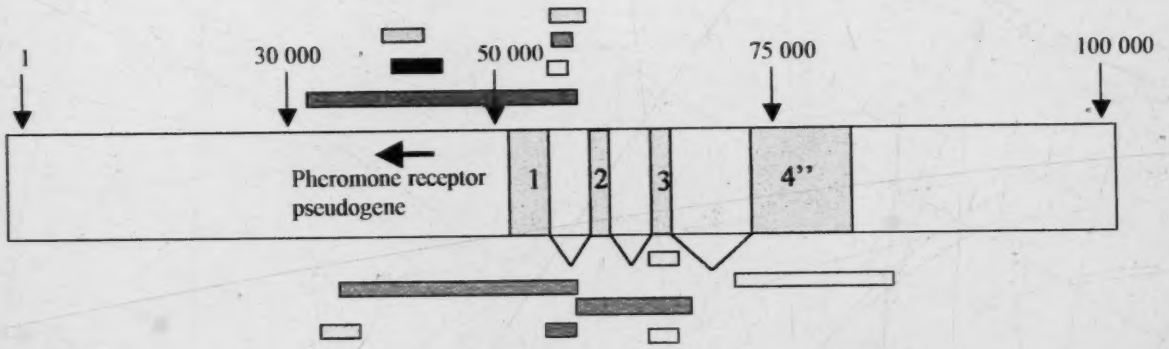


Figure 4-6: Summary diagram illustrating the complex nature of the region surrounding *CECR7* on PAC109L3. Figure is not drawn to scale. Only large regions (greater than 1 kilobase) of homology are identified. *CECR7* exons that are significantly similar to sequence elsewhere in the genome are shown, whereas repeat rich exons are not. Areas that do not show significant similarity to other chromosomes are repetitive.

cassette, and would not identify independently duplicated exons, or regions in which insertions have occurred. *CECR7* exons 2 and 3, as well as the intervening and flanking genomic sequence, appear to originate from chromosome 16p12 and were duplicated as a cassette to chromosome 22. This region on chromosome 16 was also fragmented and transposed independently to other chromosomes. *CECR7* exons 2 and 3 show similarity to a predicted gene on chromosome 16 (Loftus *et al.* 1999): the *Homo sapiens* butyryl coenzyme A synthetase I (BUCSI) (NM_052956) (annotated on clone AC003034) (Figure 4-7), and exon 3 shows similarity to the *Homo sapiens* gene similar to acetyl coenzyme A synthetase 3 (NM_058743). These chromosome 16 genes are 60% identical at the protein level and appear to have resulted from an inverted duplication. *CECR7* exons 2 and 3 show 86% and 93% identity to exons 4 and 6, respectively, of *BUCSI*. *CECR7* exon 3 shows 85% identity to NM_058742, while *CECR7* exon 2 shows no detectable identity to this gene using BLAST analysis. Interestingly, in *BUCSI*, the *CECR7*-like exons are separated by an intervening exon that is also present in *CECR7* genomic sequence (Figure 4-7); however, a splice site mutation has resulted in the loss of the exon from the *CECR7* transcript. Loss of this exon causes a frame shift in the predicted *CECR7* ORF compared to the chromosome 16 genes. Thus, *CECR7* exons 2 and 3, as well as the intervening and flanking genomic sequence, appear to originate from chromosome 16p12 (specifically from *BUCSI*, with 84.8% identity), while exon 3 and flanking sequences have also been transposed to other chromosomes. Exons 2 and 3 also show homology to clone AC025829 (also referred to as NT_033273), which is labelled as mapping to chromosome 15. However, BLAST analysis of this clone shows near 100% identity over its entire length to chromosome 16 clones (AC003034, AC020926 an

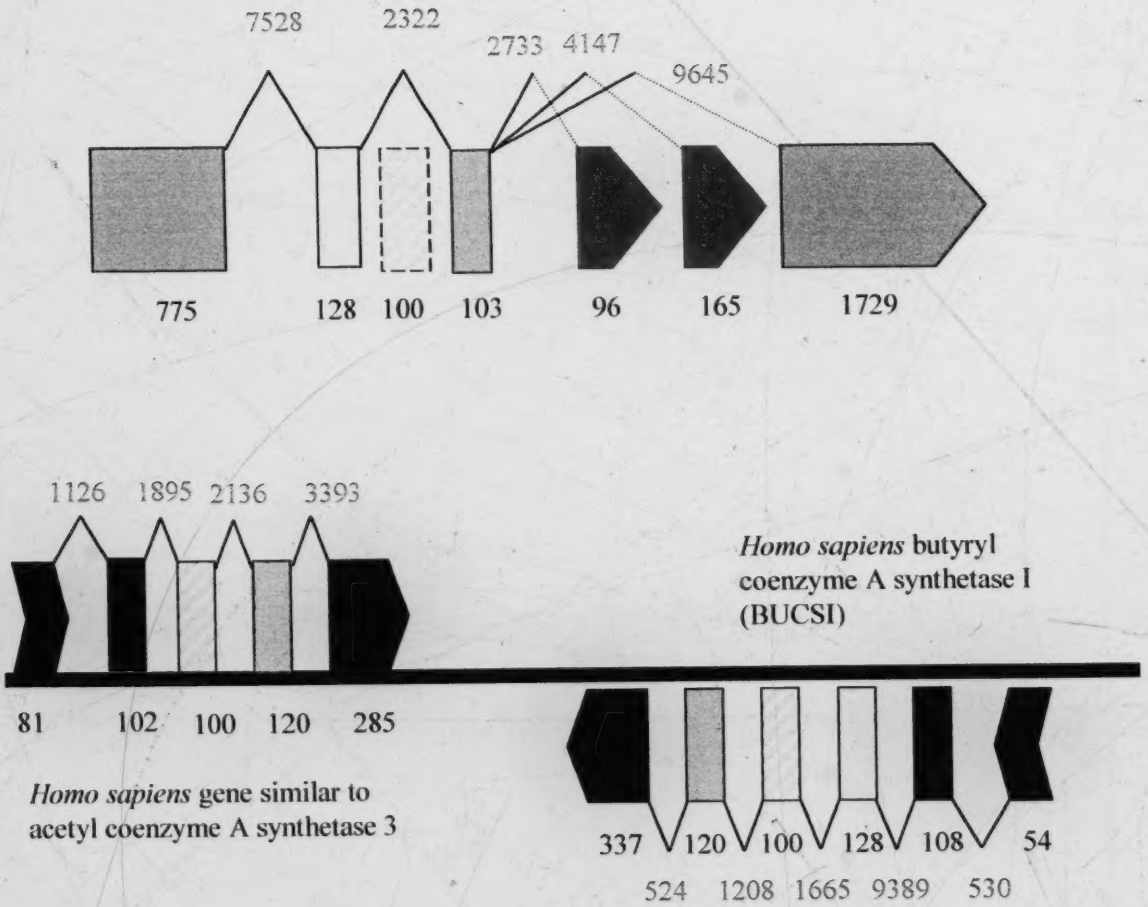


Figure 4-7: Diagram illustrating the two coenzyme A genes on chromosome 16 and the similarity of the structure of the genes to *CECR7*. *CECR7* (shown above the two chromosome 16 genes) exons 2 and 3 were duplicated as a cassette and show 86% and 93% identity to *BUCSI*. The *BUCSI* gene has an intervening exon (striped exon) between *CECR7* exons 2 and 3 that conserves the reading frame. This exon is present in the chromosome 22 genomic sequence (and is part the acetyl coenzyme A synthetase 3 gene), but is not part of the *CECR7* transcription unit. Exon sizes are labeled in black and intron sizes are labeled in blue. Similarity between exons of the various putative genes is indicated by the colouring of the exons. Exon 2 of *CECR7* is represented by the yellow shading and exon 3 is represented by the green shading.

AC004381) and genes (the coenzyme A synthetase genes described above, FLJ20274 and MACS) and nothing mapping independently to chromosome 15, suggesting that AC025829 has been mislabelled and actually maps to chromosome 16.

Alternate exons 4 and 4' (Figure 4-1) are rich in repeats, which prevent extensive sequence comparison. Exon 4'', however, is similar to sequence on chromosome 12p12 (AC009533) and 12p13.3 (AC005841). The former includes exons 4, 4' and 4'', as well as the intervening genomic sequence adjacent to the chromosome 16 duplicon. However, the similarity at 12p13.3 lies only within exon 4'' of *CECR7*, excluding (approximately) 700 nucleotides at the 5' end of the exon (Figure 4-8). A human EST (1461704, accession number AA884368) that maps to 12p12 (accession number AC009533) and shows similarity to *CECR7* exon 4''was sequenced, suggesting that exon 4'' is part of a gene being expressed from chromosome 12. This EST was derived from a pool of three normalized cDNA libraries, containing equal amounts of fetal lung, B-cell, and testis tissues. A gene could not be predicted by GENSCAN because the EST mapped very close to one end of the BAC clone, which did not overlap with any other known clone. Therefore, the sequence could not be further extended. 5' RACE was also performed to further elucidate the structure of this putative gene. Although several short products were obtained, they simply confirmed the structure already determined from the sequence of the EST. The confirmed structure of the putative chromosome 12p12 gene is illustrated in Figure 4-9.

4.4 The Evolution of *CECR7*

The formation of the pericentromeric duplicons observed in humans appears to be specific to the Great Apes, as shown by polymerase chain reaction (PCR) and

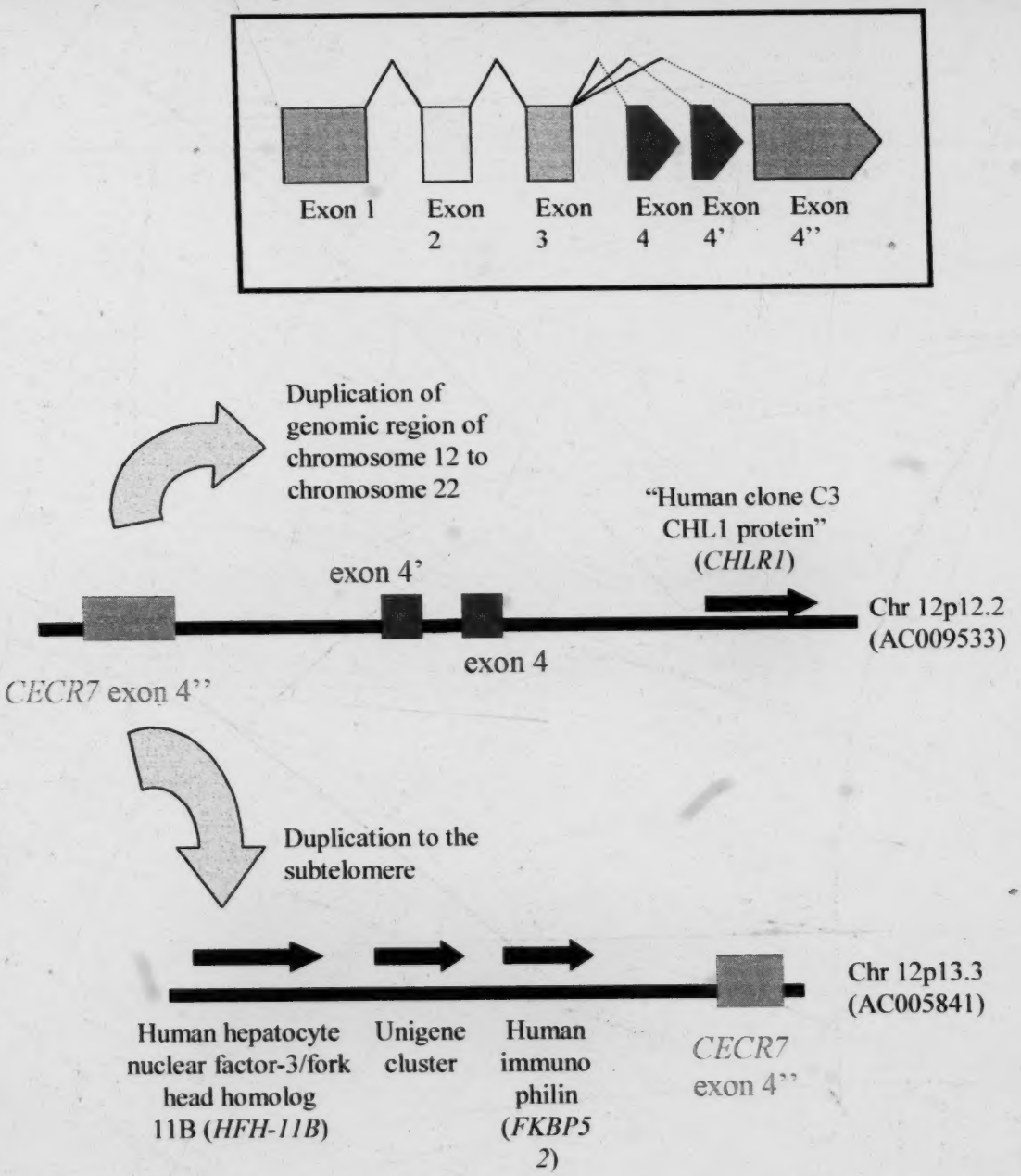


Figure 4-8: Similarity of *CECR7* exons 4, 4', and 4'' to chromosome 12. The structure of *CECR7* is depicted at the top of the figure. *CECR7* exon and intron sizes are labeled in Figure 4-7. The various colouring of the exons represents similar sequence on different chromosomes. Intronic sequence between exons 4, 4', and 4'' shows approximately 84% similarity between chromosomes 22 and 12p12.2. Figure is not drawn to scale.

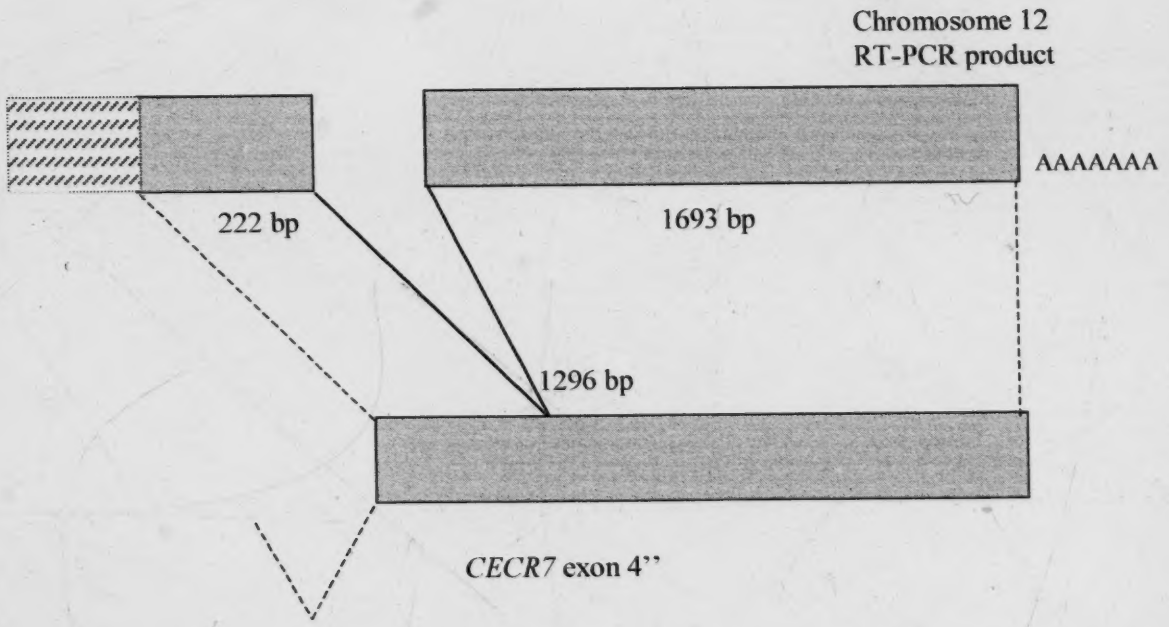


Figure 4-9: Verified structure of the chromosome 12 transcript from which *CECR7* exon 4'' likely originated. The structure was elucidated by RT-PCR, 5' RACE, and by sequencing the human EST 1461704. The putative exon and intron sizes are denoted below the gene structure. The hatched markings at the 5' end of the gene indicate the unknown upstream sequence, which cannot be determined due to a contig break. *CECR7* exon 4'' is illustrated below the chromosome 12 transcript, indicating exon 4'' shows similarity over the entire genomic region, including the putative intron of the chromosome 12 transcript.

fluorescent *in situ* hybridization (FISH) (Eichler *et al.* 1996, 1997; Horvath *et al.* 2000 a, b). However, some variation in the timing of the origin of specific duplicons has been noted. PCR was employed in various primate species to determine the evolutionary point at which *CECR7* was formed. Due to the significant sequence similarity between each *CECR7* exon and numerous other chromosomal locations, PCR was performed across the duplicon boundaries (Figure 4-1), which should only exist within *CECR7* based on monochromosomal hybrid panel results (Figure 4-3). Primers were designed to flank the two human duplicon boundaries. To minimize the effect that sequence mismatches between the various primate species had on primer annealing, low annealing temperatures were used in the PCR reactions. In addition, some primers were designed with inosine as the last nucleotide of each primer at the 3' end (since inosine will pair with all bases except guanine). These primers were designed to enhance binding to the primate DNA, which was presumed to have diverged slightly. Figure 4-10 shows the PCR results obtained across the intron 1 duplicon boundary. The primers Intron 1 F3 and R3 (Table 4-1) were used in this PCR. Results for the intron 3 duplication boundary were obtained using the primers Intron 3 F2 and Intron 3 R3 and Intron 3 F1 and Intron 3 R1 (Table 4-1) and are shown in Figure 4-11. A comparison of the intron boundary sequences from the various great ape species for both duplicon boundaries can be found in Figures 4-12 and 4-13. Results demonstrate the highly conserved nature of the introns. Over time, the sequences have not diverged a great deal. For both boundaries, PCR products (using one set of primers for each) were identified in human, chimpanzee, gorilla and orangutan. Sequencing of the products revealed over the combined total of 910bp a divergence of 0.44% between human and chimpanzee, 1.32% between human and gorilla, and 2.3% between human and orangutan, similar to recently estimated

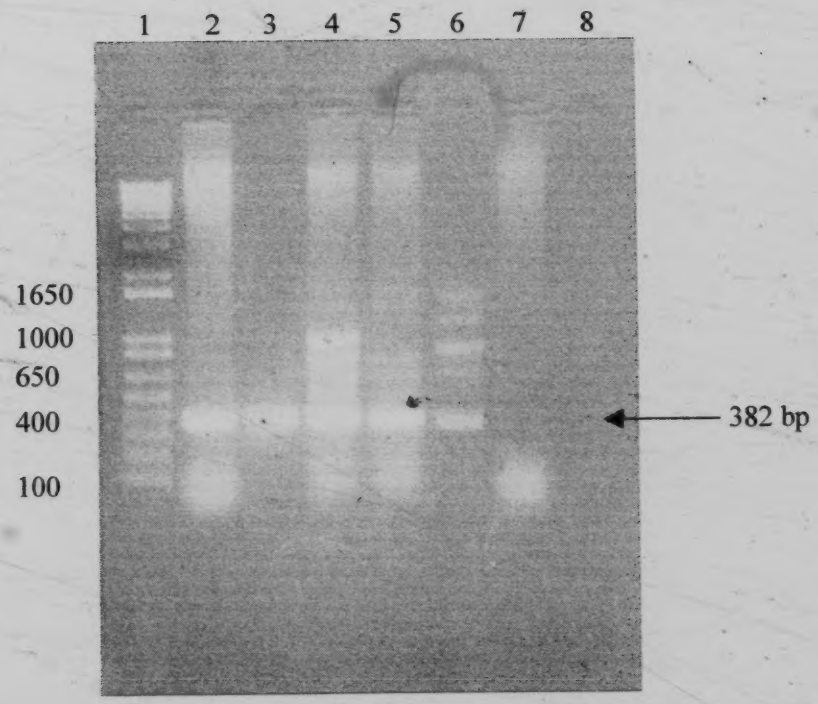


Figure 4-10: PCR products spanning the intron 1 duplicon boundary. Results indicate the intron 1 duplication boundary was formed before the divergence of orangutan, but after the divergence of macaque. Lane 1: 1kb+ ladder, Lane 2: Human Genomic DNA, Lane 3: PAC109L3 DNA, Lane 4: Chimpanzee DNA, Lane 5: Gorilla DNA, Lane 6: Orangutan DNA, Lane 7: Macaque DNA, Lane 8: Negative Control. Selected ladder sizes are labeled on the left side of the figure.

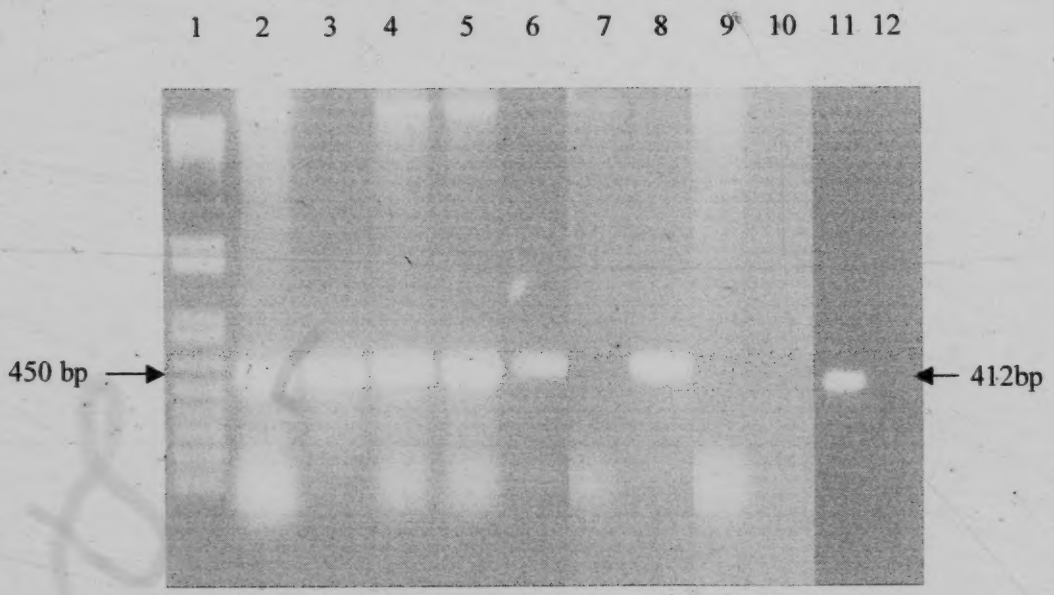


Figure 4-11: PCR results spanning the intron 3 duplicon boundary. Unlike the PCR results obtained for intron 1, results indicate that the duplication occurred before the divergence of macaque from the Great Ape lineage. The first ten lanes are results obtained using primers Intron 3 F2 and Intron 3R3. Although a PCR product was not obtained with primers Itron 3 F2 and Intron 3 R3, a product was obtained in macaque using primers Intron 3 F1 and Intron 3 R1, as shown in lane 11. Lane 1: 1 kb+ ladder, Lane 2: human genomic DNA (0.1 μg), Lane 3: PAC109L3 DNA (control clone from chromosome 22 contig), Lane 4: Chimpanzee DNA (0.1 μg), Lane 5: Gorilla DNA (0.1 μg), Lane 6: Orangutan DNA (0.1 μg), Lane 7: Macaque DNA (0.1 μg), Lane 8: Orangutan DNA (0.2 μg), Lane 9: Macaque DNA (0.2 μg), Lane 10: water control, Lane 11: Macaque DNA (0.1 μg), Lane 12: water control.

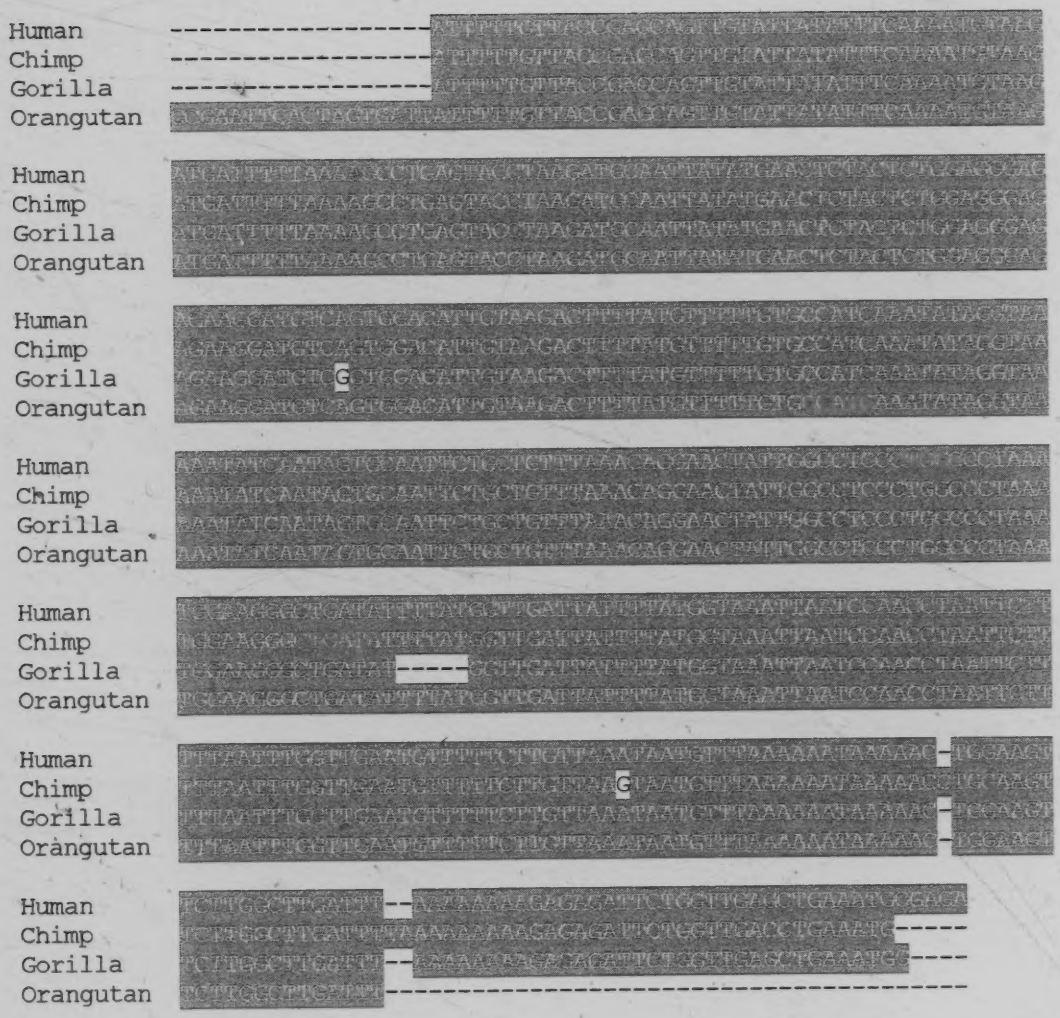


Figure 4-12: Alignment of human and primate intron 1 duplicon boundary sequences.

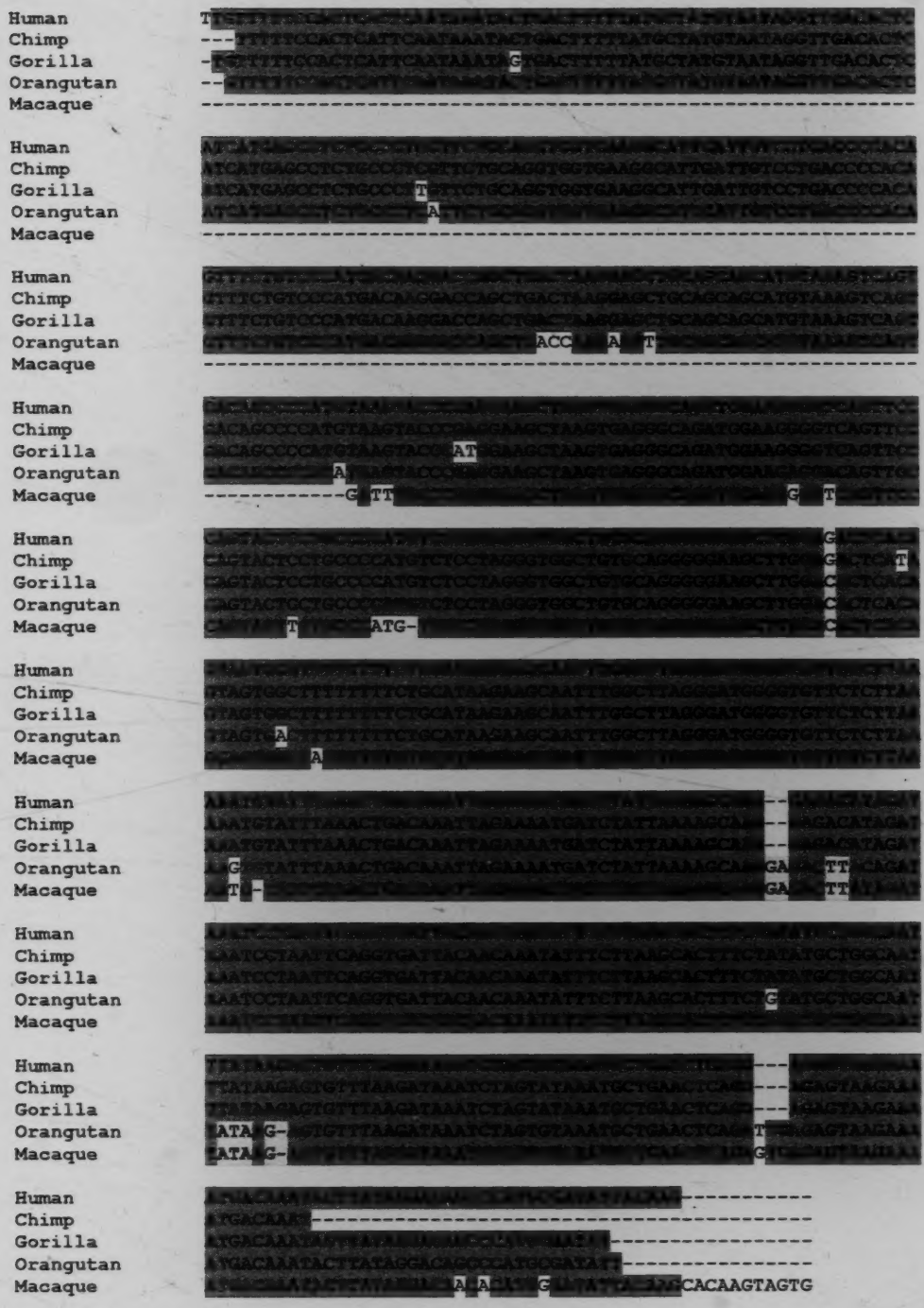


Figure 4-13: Alignment of the human and primate intron 3 duplicon boundary sequences. Nucleotides shaded in blue show conservation between macaque and orangutan, but not the other species.

divergence rates of 1.24%, 1.62% and 3.08% respectively (Chen and Li, 2001). For the intron 1 boundary in macaque, PCR products were not obtained using the Intron 1 F2/Intron 1 R3 primer set (Figure 4-10, Table 4-1). A second primer set (Intron 1 F3/Intron 1 R3) produced a PCR product, but upon sequencing the PCR product was shown to be non-specific. The intron 1 boundary PCR product from macaque was cloned and sequenced, and each clone showed similarity to different regions of the genome, but not to human chromosome 22. The PCR results suggest that the intron 1 duplicon boundary arose in the line leading to the Great Apes, before the divergence of orangutan. However, mismatches may account for the lack of macaque amplification at the intron 1 boundary. For the intron 3 boundary, four primer sets were used to test for the presence or absence of the boundary in macaque. Two of these four primer sets gave sequence-verified bands in macaque. Therefore, at least the second and third *CECR7* duplicons were formed as a unit before the separation of macaque.

The presence of *CECR7* (at the DNA level) in the Great Apes leads to the question of whether transcription occurs in the various primate species, as well. Reverse-transcription (RT)-PCR was performed on RNA derived from various human and great ape tissues to determine tissue-specific and species-specific expression of the *CECR7* transcript. The first strand cDNA synthesis reaction was performed using the oligo$(dT)_{20}$ primer. PCR was subsequently performed using a forward primer in exon 1 (Intron 1-F) and a reverse primer in exon 2 (Exon 2-R1) (Table 4-1) to ensure that all PCR products were *CECR7*-specific and not from the expressed copy of *CECR7* related ancestral genes. RT-PCR products were verified by sequencing. RT-PCR results (Figure 4-14) indicate that *CECR7* is expressed in human fibroblasts, and kidney, but not in human lymphoblasts. *CECR7* is also expressed in gorilla

fibroblasts, but not in gorilla or chimpanzee lymphoblasts, as would be expected based on the human expression pattern. RNA was not available from chimpanzee fibroblasts. Expression was not detected in orangutan fibroblasts using either the oligo$(dT)_{20}$ primer or two different gene specific primers (SAHL 11R1, SAHL 11R2, which are in *CECR7* exon 3) (Table 4-1) for the reverse transcription reaction, suggesting *CECR7* is not expressed in the orangutan (Figure 4-15) or macaque. Alternatively, the primers used may have been too diverged in sequence to anneal to the RNA, or to the cDNA. Figure 4-16 is an alignment of the sequence of the human kidney RT-PCR product and the gorilla fibroblast product.

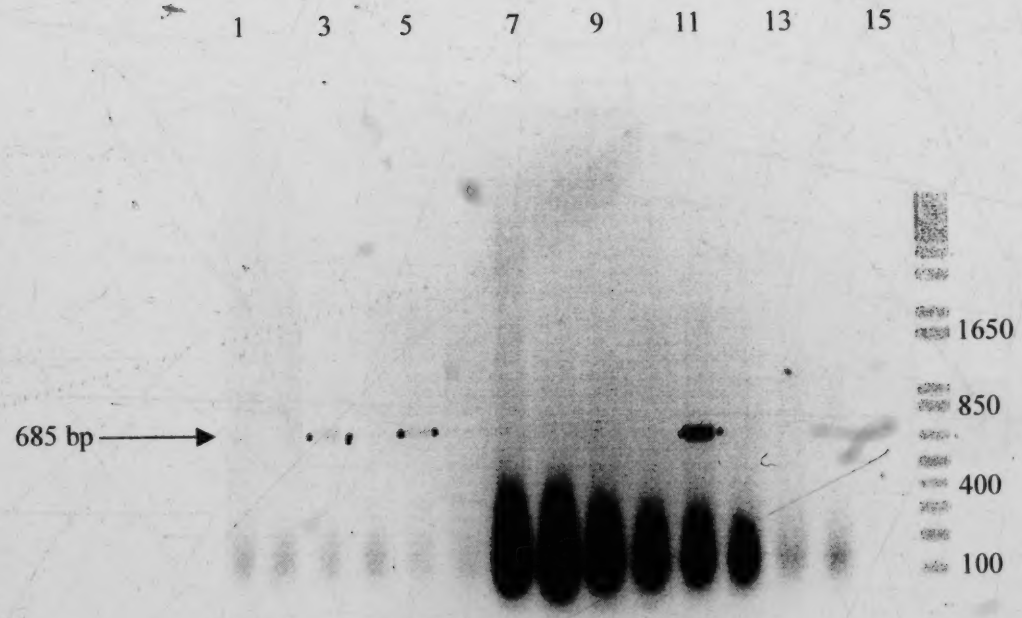


Figure 4-14: RT-PCR results indicating *CECR7* is expressed in gorilla fibroblasts. Bands are marked by dots on either side on this reverse image of an ethidium bromide-stained agarose gel. Results indicate *CECR7* is expressed in human fibroblasts and kidney, but not in human, chimpanzee or gorilla lymphoblasts. *CECR7* is also not expressed in macaque fibroblasts. The reverse transcription reaction was performed using the oligo$(dT)_{20}$ primer and the PCR reaction was performed using the primers Intron 1-F and Exon 2 R1. Lanes labeled contain the RT-PCR reaction. The lane following the labeled lane is the negative control, which contained RNA, but no reverse transcriptase. Lane 1: Human lymphoblast, Lane 3: Human fibroblast, Lane 5: Human kidney, Lane 7: Chimpanzee lymphoblast, Lane 9: Gorilla lymphoblast, Lane 11: Gorilla fibroblast, Lane 13: Macaque Fibroblast, Lane 15: PCR negative control, Lane 16: 1kb+ ladder. Selected ladder sizes are labeled on the right side of the gel.

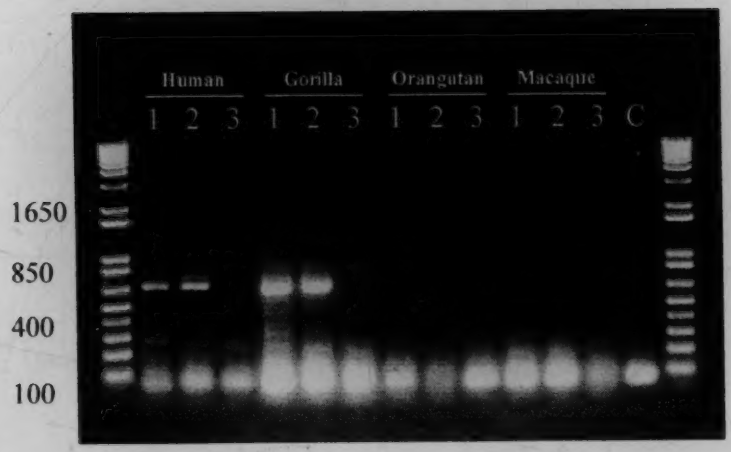


Figure 4-15: RT-PCR results demonstrating *CECR7* expression in human and gorilla fibroblasts, but not in orangutan or macaque fibroblasts (Bridgland *et al.* in press, SAHL 11R2 reactions performed by Melanie Kardel). Two different reverse primers (SAHL-R1 and SAHL-R2) were used for cDNA synthesis to minimize the chance of false negatives. Lanes are marked as follows: 1. RT reaction performed with the R1 primer. 2. RT reaction performed with the R2 primer. 3. Negative RT control. C. water control. Selected ladder sizes are labeled on the left side of the gel.

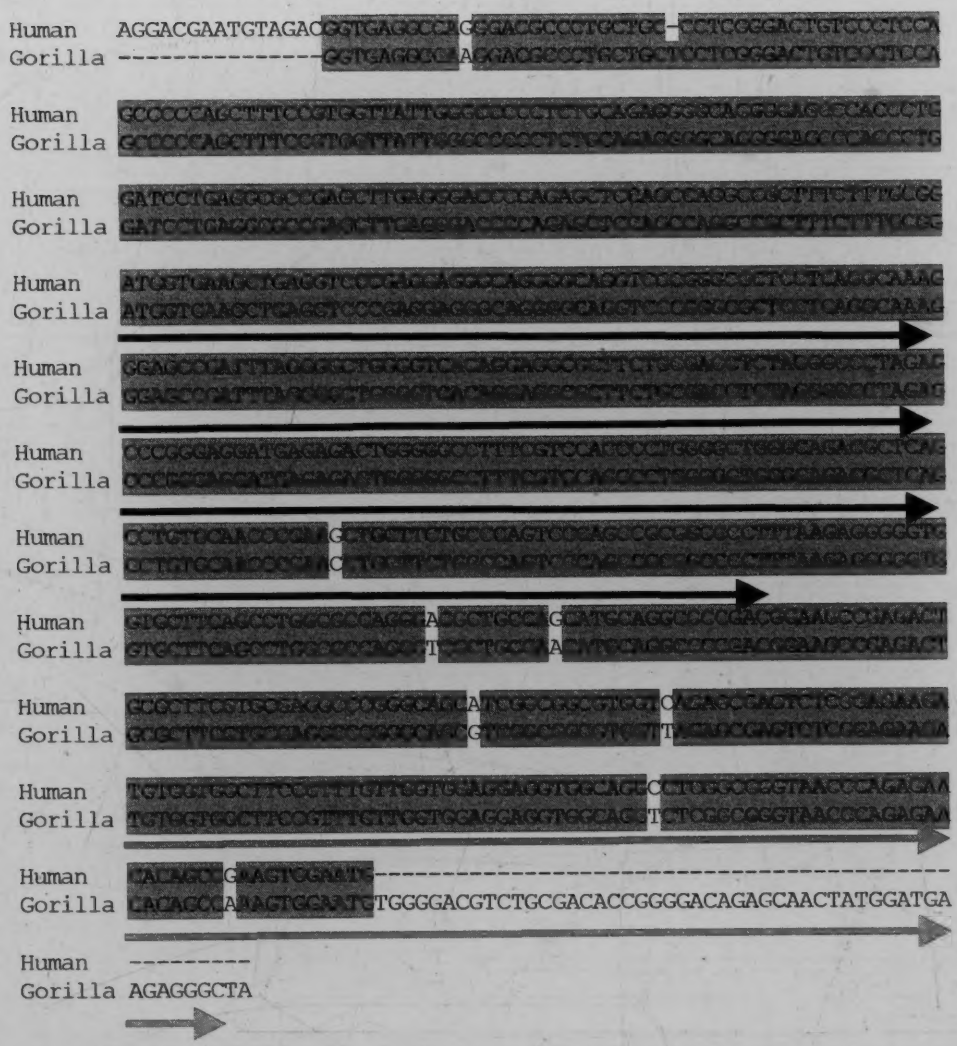


Figure 4-16: Alignment of *CECR7* RT-PCR product from human kidney and gorilla fibroblast. Although there are several changes between human and gorilla in the sequence, these sequence differences do not alter putative open reading frames. The open reading frames are marked below the sequence with arrows. ORF-A (frame 0) is marked in black and ORF-C (frame +2) is marked with blue arrows.

4.5 Cloning of *CECR8*

CECR8 was identified by 3' ESTs of two cDNAs (1840206 and 1840041), which showed 98-100% identity to chromosome 22 when sequenced. Both cDNAs originate from testis cDNA libraries. These ESTs identified two exons that are 83% and 89% identical to sequence on chromosome Y (accession number AC006040). The cDNA 1840206 may represent an intronless alternative 3' end of the gene (one of the *CECR8* transcripts may be a single exon). After the two cDNAs were sequenced, the GENSCAN gene prediction program was utilized to predict gene sequence upstream on PAC8708 (Figure 1-2). However, due to a contig break, no 5' exons could be predicted. To further extend the gene sequence, 5'RACE was performed using various adult and fetal cDNA libraries. *CECR8* products were obtained from both fetal brain and adult testis RACE libraries. Although these PCR products were chromosome 22-specific, little novel sequence was obtained. RACE products from testis further extended the sequence of putative exon 1 (identified by hEST 1840041) by approximately 25 nucleotides. Other chromosome 22-specific RACE products ran into the intron of *CECR8*, but were too small to determine whether there is an intronless alternative transcript or if splicing to exon 1 occurs.

4.6 Alternative Splicing Identified in *CECR8*

RT-PCR was used to further elucidate the structure of *CECR8*. Since testis RNA was not available, a testis cDNA library (prepared with an oligodT_{20} primer) was used in place of the reverse transcription reaction. A forward primer in putative exon 1 (*CECR8* 3F) and a reverse primer in exon 2 (*CECR8* 2R) (Table 4-1) were used for the PCR. RT-PCR products were very faint and not easily identifiable on a gel. Therefore, these bands were extracted from the gel, diluted in water and the same

primers were then used in a secondary PCR reaction. Four faint bands were identified in the second round of PCR and were subsequently subcloned and sequenced. These bands represented four alternative transcripts. *CECR8* alternative splicing is illustrated in Figure 4-17.

4.7 Identification of *CECR8*-like sequence

BLAST analysis identified several regions of the human genome that show extensive similarity to *CECR8* (Figure 4-17). Although this putative gene likely arose from a duplication and transposition event, *CECR8* is not as complicated in its evolution as is *CECR7*. *CECR8* is highly similar to sequence on the Y chromosome. BLAST searches performed on the non-redundant database (nr) and the high throughput genomic sequence (htgs) database identified several Y chromosome clones that contain *CECR8*-like sequence. For example, *CECR8* is 89% similar to sequence on a Y BAC clone (AF114156), which contains the Y-linked zinc finger (ZFY) gene. Another Y chromosome clone, AC006040, also contains sequence similar to *CECR8*. However, on this clone there appears to be splicing (ie. an intron exists) between putative *CECR8* exons (as opposed to the sequence on AF114156, which appears to be a duplication without the intron, perhaps from the integration of an mRNA), as illustrated in Figure 4-18. BLAST searches identified two exons on AC006040 that are approximately 85% similar to exon 1 and the largest *CECR8* exon 2. A search of the human EST database identified many ESTs that are 98%-100% identical to several of these Y chromosome clones, making it difficult to determine the evolutionary origin of the chromosome 22 copy of *CECR8*. *CECR8* is also 88% similar to the *Homo sapiens* heat shock transcription factor 2-like mRNA

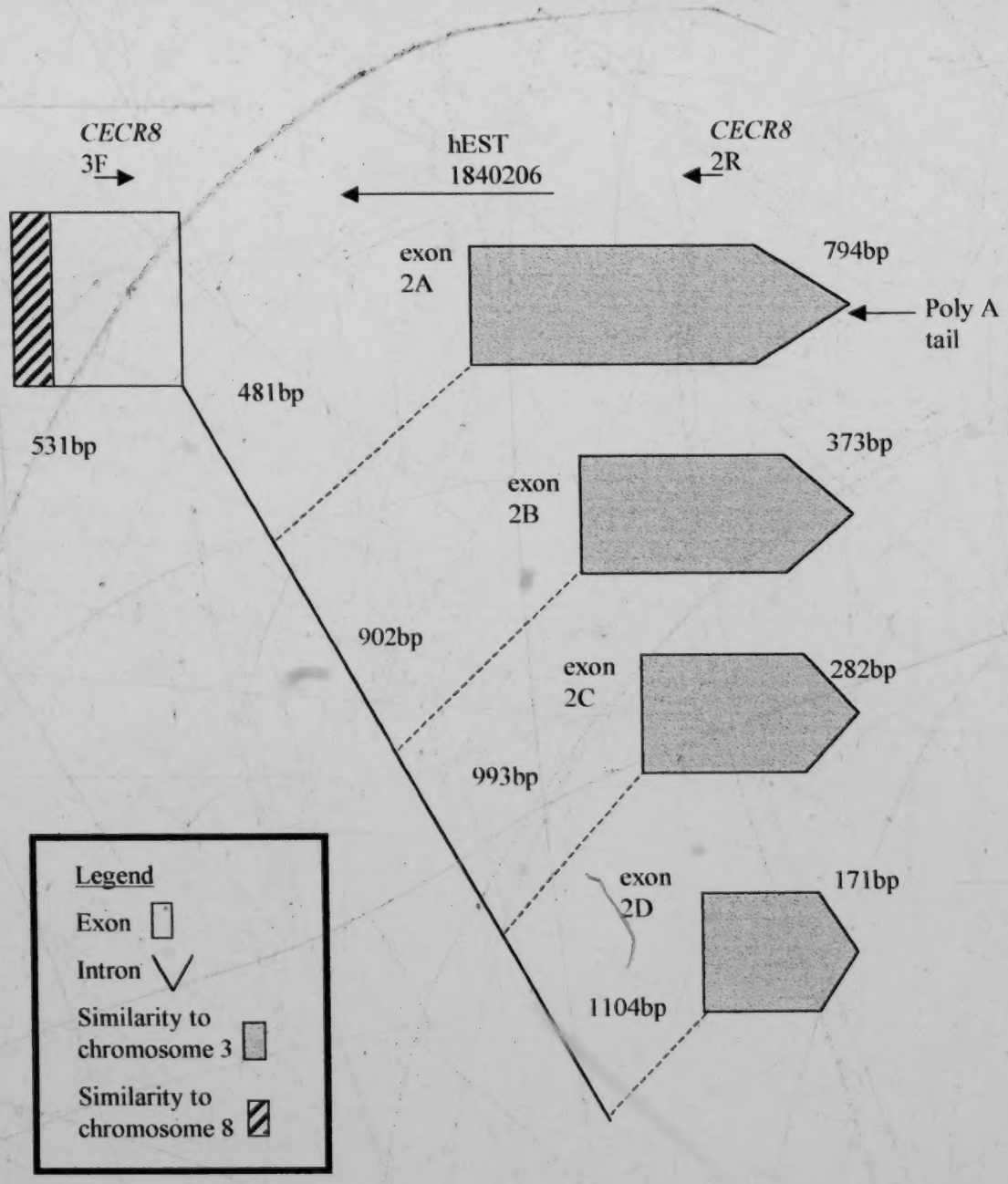


Figure 4-17: Alternative splicing observed in *CECR8*. All four transcripts share the same exon 1, but splice to different downstream regions to form four different versions of exon 2. All four transcripts share the same polyA tail. Primers used in RT-PCR are labeled on the diagram. Exon 2B represents hEST 1840041, whereas hEST 1840206 is labeled on the diagram. *CECR8* is 89% similar to chromosome Y over its entirety.

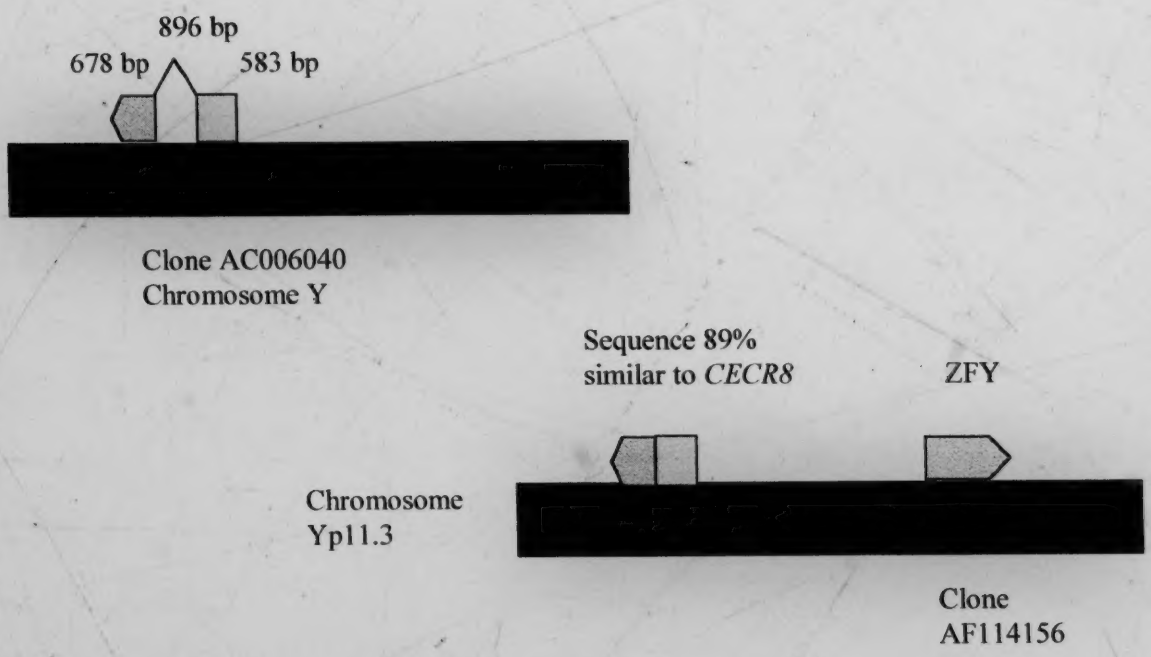


Figure 4-18: Similarity of *CECR8* to sequence on the Y chromosome. *CECR8* is 89% similar to sequence on the Y chromosome. Based on BLAST analysis, *CECR8* is similar to many Y chromosome clones, suggesting *CECR8* was not only duplicated to the pericentromere of chromosome 22, but also other locations on the human Y chromosome. Two of the clones on which *CECR8*-like sequence is found are depicted in the above diagram. On AC06040, there is evidence of splicing (ie. there are two exons similar to those of *CECR8* and an intervening intron) in the *CECR8*-like sequence, suggesting this was the origin of the duplication. *CECR8*-like sequence is also found on AF114156 in close proximity to the Y-linked zinc finger gene (ZFY). There is no intron on AF114156.

(XM_044000.1), which is also found on the Y chromosome. Recently, a testis cDNA from an adult male mouse was identified. This cDNA is 79% similar to *CECR8* over the second half of exon 1 and encodes an HSF-type DNA-binding domain containing protein. As illustrated in Figure 4-17, portions of *CECR8* are also similar to other locations in the human genome. Exons 2B, 2C, and 2D are approximately 83% similar to sequence on chromosome 3 over their entirety. Exon 2A is also 83% similar to chromosome 3 sequence, but only over the region of the exon which is also common to exon 2B. The first quarter of exon 1 is 88% similar to chromosome 8 sequence.

4.8 Expression Analysis of *CECR8*

Several human adult and fetal northern blots were probed with hEST 1840041 to examine in which tissues this putative gene is expressed. Results are shown in Figure 4-19. Of all tissues examined (brain, heart, skeletal muscle, kidney, liver, placenta, and lung), a transcript of approximately 1.6 kb was identified only in adult testis. However, 5'RACE results indicate *CECR8* is also expressed in fetal brain, but likely at very low levels, explaining its absence on a Northern blot. In addition, RT-PCR and 5'RACE were attempted for *CECR8* using adult heart, kidney, and liver RNA or cDNA libraries and were continuously unsuccessful, confirming Northern results. If the band observed on the Northern blot represents the largest *CECR8* transcript, at least 0.4 kb of this putative gene still needs to be cloned.

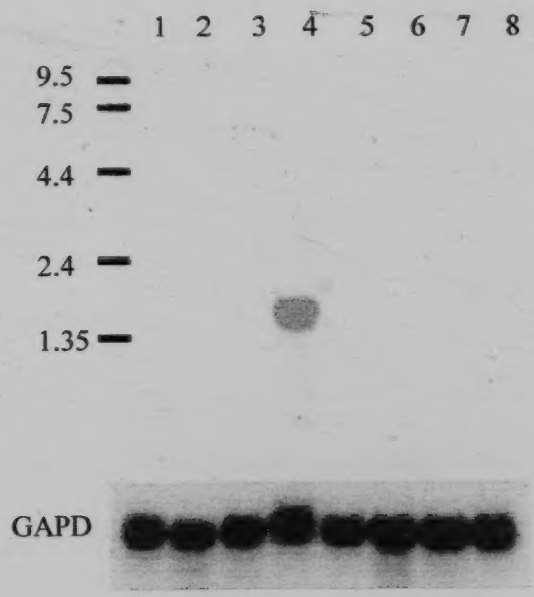


Figure 4-19: Northern blot results for *CECR8*. Several northern blots were probed with the human EST 1840041. Expression of *CECR8* could only be detected in testis, as shown above. The control probe GAPD was also probed on this blot and shows even loading of the lanes. Lane 1: spleen, Lane 2: thymus, Lane 3: prostate, Lane 4: testis, Lane 5: Ovary, Lane 6: small intestine, Lane 7: colon, Lane 8: peripheral blood leukocytes.

Chapter 5: Discussion

5.1 Significance of Proposed Origins of *CECR7*

Until recently, pericentromeric regions were believed to be junkyards for genomic DNA fragments. The extensive degree of duplication noted in pericentromeres made it extremely difficult to sequence these regions and to assign them a location in the human genome. Despite these drawbacks, several groups have continued to concentrate their research efforts on pericentromeres (Eichler *et al.* 1996, 1997; Jackson *et al.* 1999; Horvath *et al.* 2000a/2000b). For example, Eichler has studied the evolution of pericentromeres, focusing his research on the determination of the time point at which the duplications arose, as well as tracing the origin and spread of individual duplicons.

Although this field has become increasingly popular, to date most of the data has involved either the identification of genes with novel expression patterns that arose via duplication (Iyer *et al.* 1996) or the identification of the presence of these pericentromeric duplicons in some primate species and their absence in others (Regnier *et al.* 1997; Jackson *et al.* 1999). In contrast, *CECR7* is the first reported example of the formation of a chimeric transcription unit in a pericentromeric region (Footz *et al.* 2001, Bridgland *et al.*, in press). This putative gene creates a novel opportunity to further unravel the process of genome evolution, as the origins of this transcription unit have yet to be obscured by noise caused by genetic mutation.

Determining the original location of *CECR7* exons, which is possible due to its recent creation, may lead to interesting questions regarding possible functions of this transcription unit. For example, since *CECR7* appears to have arisen from the duplication of at least two (or perhaps three) genes, it would be interesting to compare the function of these genes to that of *CECR7*. Perhaps *CECR7* will have a function

that is similar to one of the known genes. Since none of the four putative ORFs substantially span more than one duplicon, it is unlikely that *CECR7* will have a novel function through the combination of ORFs. However, a novel function may arise due to differences between regulatory elements surrounding *CECR7* and the exons of *BUCS1* or the putative chromosome 12 gene. This possibility also raises the question of how genes acquire function. At the very least, tracing the origins and the spread of the duplicons of this putative gene will lead to a better understanding of genome evolution.

Furthermore, it is not only important to study the origin of *CECR7,* but it is essential to analyse the spread of the duplicons following the initial duplication event to further understand genome evolution. Although at this point it is not known why pericentromeres are hotspots for the integration of duplicated fragments of DNA, the study of the spread of duplicons may reveal the answer. Eichler *et al.* (1999) identified a CAGGG repeat unit in pericentromeres that was associated the duplication and integration of five known genes in pericentromeric regions (see Section 2.4). However, this CAGGG repeat is not present in the pericentromeric region surrounding *CECR7,* suggesting at least one other mechanism may be responsible for this process of duplication and eventual integration into pericentromeres.

Determining the spread of *CECR7* exons proved to be problematic. Since the genome has not been sequenced in its entirety as of yet, BLAST analysis can identify most, but not all, of the paralogous *CECR7* locations. The origins of the *CECR7* exons should be easily identifiable by BLAST analysis since they are presumed to be located in the gene-rich chromosomal arms (which are fully sequenced), unlike pericentromeric duplicons, which are in regions that are difficult to sequence and

assign to a locus given the repetitive nature of these areas. Therefore, PCR was performed on the monochromosomal hybrid panel to confirm, and further identify, chromosomes on which *CECR7*-like sequence is present. The only exon for which the PCR yielded reliable results was exon 1. Monochromosomal hybrid panel PCR identified exon 1-like sequence on two chromosomes (5 and 19) that was not previously identified by BLAST analysis. However, PCR using monochromosomal hybrid DNA did not amplify known exon 1-like sequence on chromosomes 10, 14, 18, and 22. There are several possible explanations for this. With the exception of the chromosome 22q11 pericentromeric region, the sequence of the primer may have been too diverged for proper annealing during PCR. For instance, a change in one nucleotide at the 3' end of the primer may greatly decrease primer annealing, explaining the absence of results. Moreover, monochromosomal hybrid cell lines are known for rearrangements. Perhaps a rearrangement occurred in the pericentromere of chromosome 22 (or on any other chromosome) that resulted in a region too large to amplify by PCR. This would explain the lack of a PCR product using chromosome 22 specific primers with the monochromosomal 22 hybrid. Two bands were evident on chromosome 3 (Figure 4-4). The larger band, which was barely visible, was sequenced, but sequence was too poor to confirm whether it was a paralogous location of *CECR7* exon 1. This band was the same size as all other bands confirmed by sequence to be paralogues of exon 1, suggesting it may be similar to *CECR7* exon 1, as well. The smaller band was not sequenced. Since the similarity of chromosome 3 sequence to *CECR7* could not be determined, it is not possible to postulate whether the origin of *CECR7* exon 1 is on chromosome 3. In addition, there was a chromosome 5 band (that was larger than expected) that was sequenced, but sequence reads of PCR products were poor (they contained many ambiguously-called

nucleotides) and significant sequence similarity to elsewhere in the genome could not be elucidated.

As has been previously suggested in the Results Section (see pg. 49), the possible origin of exon 1 may be on chromosome 14, or elsewhere on chromosome 22, based on the predicted age of the duplication expected from the degree of divergence. A genomic duplication of this region of DNA obviously occurred, and then likely spread to a pericentromere. Subsequent smaller duplications then spread to other pericentromeres, including chromosome 22q. Since the hypothetical gene supported by mRNA AK056135 and the sequence surrounding it are 100% identical to both chromosome 14 and 22, there was likely a very recent duplication of the region from the original location to the arm of the other chromosome (ie. 14 to 22 or vice versa), in addition to the duplication event that seeded the pericentromeric duplications. Similarly, putative exons of this hypothetical gene have also been duplicated to other regions of chromosome 14 and 22, suggesting this area of the genome containing *CECR7* exon 1-like sequence (which is part of the hypothetical gene that has been duplicated several times) is a hotspot for duplication events.

CECR7 exons 2 and 3 were duplicated from chromosome 16, a region of the genome that also appears to be susceptible to duplication events. Exons 2 and 3 show similarity to the predicted exons of the *BUCS1* (NM_052956) gene on chromosome 16 (annotated on clone AC003034), and were duplicated as a cassette from chromosome 16 to chromosome 22 (see Figure 4-7). Exon 3 also shows similarity to a predicted exon of the *Homo sapiens* gene similar to acetyl coenzyme A synthetase 3 (NM_058743) (also annotated on clone AC003034). Although *CECR7* exon 2 is only similar to sequence on chromosome 16, exon 3 is similar to sequence on several different chromosomes, as well as to many other locations on chromosome 16, in

addition to *BUCS1* and the *Homo sapiens* gene similar to acetyl coenzyme A synthetase 3. Again, sequence that was duplicated to the pericentromere of chromosome 22 and was integrated into the *CECR7* transcription unit originates in an area of the genome that appears to be prone to duplication.

The predicted *BUCS1* transcript includes an intervening exon between *CECR7* exons 2 and 3. Although the *CECR7* mRNA does not contain this putative exon (due to a splice site mutation), the genomic sequence on chromosome 22 does. The lack of this exon in the chromosome 22 transcript results in a frame shift. The question of whether *CECR7* is functional will be further addressed; however, it is interesting to ponder whether the intervening exon was lost due to random chance or if the *CECR7* transcription unit with the exon was selected against. If the duplication of the chromosome 16 cassette to chromosome 22 resulted in a functional protein similar to the *BUCS1* gene, perhaps this gene was dosage sensitive and was selected against due to the detrimental effects of excess protein. This would, in turn, lead to selection in favour of the transcript with the splice site mutation, which alters the reading frame and abolishes *BUCS1*-like activity. However, since these three exons do not comprise the majority of the *BUCS1* ORF (in fact, they account for only a very small portion of the protein), it is unlikely that the three exons produce a protein that functions in a similar manner to *BUCS1*. Therefore, it is more plausible that the splice site mutation occurred by chance. Since an extra copy was not required, mutations accumulated.

The duplication of the genomic region containing *CECR7* exons 4, 4', and 4'' originated at chromosome 12p12. Exons 4 and 4' are rich in repetitive sequence whereas exon 4'' is similar to sequence at 12p12 (84.5% identity to clone AC009533), as well as 12p13.3 (85.8% identity to clone AC005841). *CECR7* intron 3 sequence is similar to chromosome 12p12, but not to 12p13.3. This suggests that

exon 4'' originated at 12p12, a region of euchromatin, and was partially duplicated to the subtelomeric region of 12p13.3. Duplication and transposition of the larger cassette to chromosome 22q11 occurred independently. Conversely, the duplication may have originally been transposed to 22q11 followed by partial exchange with 12p13.3 (Figure 5-1). Exchanges between pericentromeric and subtelomeric regions have been previously reported (Eichler 1998; Trask *et al.* 1998). For example, there are a large number of olfactory receptor (OR) genes in the human genome. In fact, greater than 25 OR sequences reside in the human genome, many of which are located at various telomeres (Trask *et al.* 1998).

The putative gene from which exon 4'' was derived has only been partially cloned. It would be interesting to determine whether this putative gene does in fact encode a protein and to determine the function of that protein. The known sequence of the putative chromosome 12 gene does not contain an open reading frame, likely because the region cloned is the 3' untranslated region. Although it is more likely that the location of the gene would determine the mechanism by which duplications occur, perhaps the function of the chromosome 12 gene may reveal a trend in the type of genes that undergo duplication. In addition, *CECR7* ORF-D spans the 5' end of exon 4'' (see Figure 4-1). Based on BLAST analysis, ORF-D does not show significant similarity to any known protein. This ORF is not present in the chromosome 12 gene due to the sequence differences between the two putative genes.

Whether exons 4 and 4' are part of the chromosome 12 gene is unclear due to a contig break. If they are not part of this gene, why (and how) did they become alternative 3' exons for the *CECR7* transcript? Are such events fortuitous? Aside from splice acceptor and donor sites, other factors must play a role in the determination of the structure of a transcript. For example, certain sequences

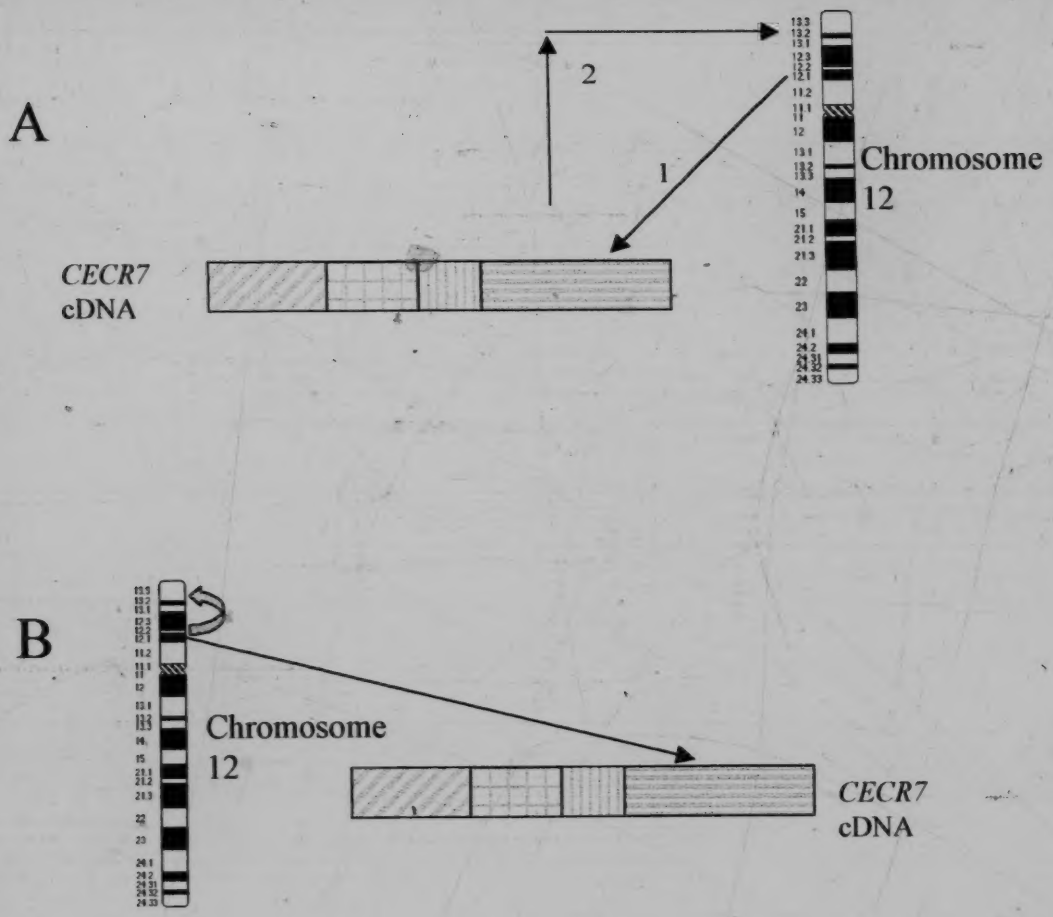


Figure 5-1: Hypothesized duplication events leading to the presence of *CECR7*-like sequence on chromosome 22, 12p13.3, and 12p12. Panel A shows the duplication of exon 4'' from 12p12 to chromosome 22 (#1), followed by partial exchange with the subtelomere of chromosome 12 (#2). Panel B illustrates partial duplication of 12p12 to the subtelomere, as well as independent duplication of 12p12 to chromosome 22q11.2.

surrounding exons have been shown to affect alternative splicing. Alternative splicing of alpha-tropomyosin depends on a mutually exclusive selection of exons 2 and 3 (Gromak and Smith, 2002). Repeat elements on both sides of exon 3 in the alpha-tropomyosin gene are required for the inhibition of exon 3 and the selection of exon 2 in the smooth muscle transcript. Mutations in these repeat elements disrupt the proper alternative splicing in smooth muscle. Exonic splicing enhancers, which enhance splicing at nearby splicing sites, have also been reported to play a role in alternative splicing (Fairbrother *et al.*, 2002). Future studies may help elucidate certain mechanisms associated with gene formation once the structure of the chromosome 12 gene is determined. Such studies may include examining the exon/intron boundaries for obvious repeat elements, or searching for exonic splicing enhancer elements similar to those previously reported.

In contrast to the other duplicons from which *CECR7* was formed, the chromosome 12 sequence does not appear to be as susceptible to duplication events. The limited dispersal of the chromosome 12 duplicon may be due to a low rate of exchange between pericentromeres. It is also possible that sequences surrounding the other regions that are highly duplicated contain a particular sequence that is important for signalling the duplication event. For example, Eichler (Eichler *et al.*, 1996) identified an interspersed CAGGG repeat sequence within 16p11.2 at the junction of the duplication of a sequence from Xq28. CAGGG repeats were then characterized (Eichler, 1999) and found to be almost exclusively distributed within pericentromeric regions of primate chromosomes. In addition, these repeat elements were found to be present in the hominoid lineage prior to duplication of the sequences with which they were associated and, most importantly, the CAGGG repeats delineated the boundaries of several recent genomic duplication events in pericentromeric regions. Therefore, it

has been proposed that the CAGGG repeat elements in the pericentromeric region may mediate the integration of duplicated segments (Eichler 1999). Repetitive tracts of three or more guanine residues have been associated with the formation of G4 DNA (Williamson 1994), which has been implicated in homologous pairing of chromosomes and recombination. Interestingly, the CAGGG repeats identified by Eichler (1999) are strikingly similar to switch-recombination signals that are involved in genomic rearrangement events and result in the recombination of the VDJ immunoglobulin regions. However, the region containing *CECR7* on chromosome 22 was examined and this repeat element was not present (Footz *et al.* 2001), suggesting at least one more mechanism exists to mediate duplication and integration events at the pericentromere. Since subtelomeric regions are also rich in duplications, it would be interesting to examine whether a universal, genome-wide, signal exists, or if a specific element mediates duplication events at subtelomeric regions, and a different element is involved in the duplication and integration at pericentromeres. G-rich DNA has been identified in switch-recombination regions, telomeres, rDNA, and now in pericentromeres, suggesting G-rich DNA may play a role in the evolution of these regions by promoting pairing and recombination of duplicated segments of DNA. However, CAGGG repeats were found to be exclusively pericentromeric, indicating that other G-rich sequences may act as preferred sites of integration for duplicated fragments of DNA. Future work may include a search for specific sequence elements that are involved in pericentromeric duplication events, but is currently limited by the availability of centromeric sequence. An approach similar to that taken by Eichler would be useful in identifying other possible sequence elements involved in the integration of duplicated DNA at pericentromeres. Once complete sequence is available, it will be more reasonable to study/track the distribution of pericentromeric

duplications and differentiate duplicons from their origins (based on their locations in pericentromeres or chromosomal arms). Computer algorithms can be utilized to identify possible G-rich repeat elements in regions surrounding duplicons. FISH could then be used to probe human and metaphase spreads to determine whether putative localization signals are specific to the pericentromere or spread throughout the genome.

The focus of this thesis is *CECR7*, which was the first reported example of a chimeric transcript formed over multiple duplicons in a pericentromeric region (Footz *et al.* 2001). A second, similar chimeric transcript within the chromosome 22q pericentromeric region has also been described (Bailey *et al.* 2002). The AK001299 transcript maps to PAC 15J16 and is thus approximately 200 kb proximal to *CECR7* (Footz *et al.*, 2001). AK001299 consists of 3 exons, each derived from a different duplicon with similarities to different chromosomes (7q36, 13q14 and 12p13 for exons I, II, and III respectively). However, unlike *CECR7*, none of the AK001299 exons have been shown to be derived from putative ancestral genes (exon III shows similarity to von Willdebrand factor exon sequence, but in the opposite orientation). Similar to *CECR7*, the AK001299 transcript shows only a small possible ORF of 98 residues. Several other fusion genes created through partial gene duplication were identified on chromosome 22 (Bailey *et al.*, 2002), but none of these were in the pericentromeric region. Transcripts NM_021574, U03891 and AW190323 all show fusions between exons from an intrachromosomal partial duplication of a gene and exons in the DNA adjacent to the duplication.

5.2 *CECR7*: Identification of the Time Point of the Duplication Events

PCR analysis indicated that the intron 1 genomic duplicon boundary was formed before the divergence of orangutan from the Great Ape lineage (Figure 5-2), similar to results reported by Regnier *et al.* (1997). This group demonstrated that the duplication and subsequent transposition events involving the Type I neurofibromatosis (*NF1*) gene, which led to the presence of at least 11 *NF1*-related sequences in the human genome, began after the divergence of macaque. Regnier *et al.* demonstrated by PCR that the ancestral *NF1* gene was present in only one copy in macaque, in contrast to the three bands obtained by PCR in gibbon. To rule out DNA mismatches being the cause of the absence of PCR products, four different primer sets were used and a single band was obtained in macaque for each primer set. The PCR band was sequenced and showed significant similarity to the functional human gene. Three different sequences were identified by PCR in gibbon (the species which diverged before orangutan and after macaque). Two of the sequences were similar to the human *NF1*-related sequences, whereas the third showed similarity to the functional human gene. To confirm these results, Regnier *et al.* performed FISH on macaque metaphase spreads and detected only one signal, which mapped to the macaque chromosome homologous to human chromosome 17. These results demonstrate that the duplication of the *NF1* gene began after the divergence of macaque (22-33 mya) from the Great Ape lineage.

Similarly, Jackson *et al.* (1999) have studied the duplication-rich pericentromeric region of chromosome 10 and have concluded that the chromosome 10 *ZNF* and D10S141 sequences were duplicated before the divergence of orangutan (15-22 mya). Analysis of a 9.75Mb YAC-based map across the centromere of human chromosome 10 indicated that duplicated ZNF (zinc finger) gene clusters were linked

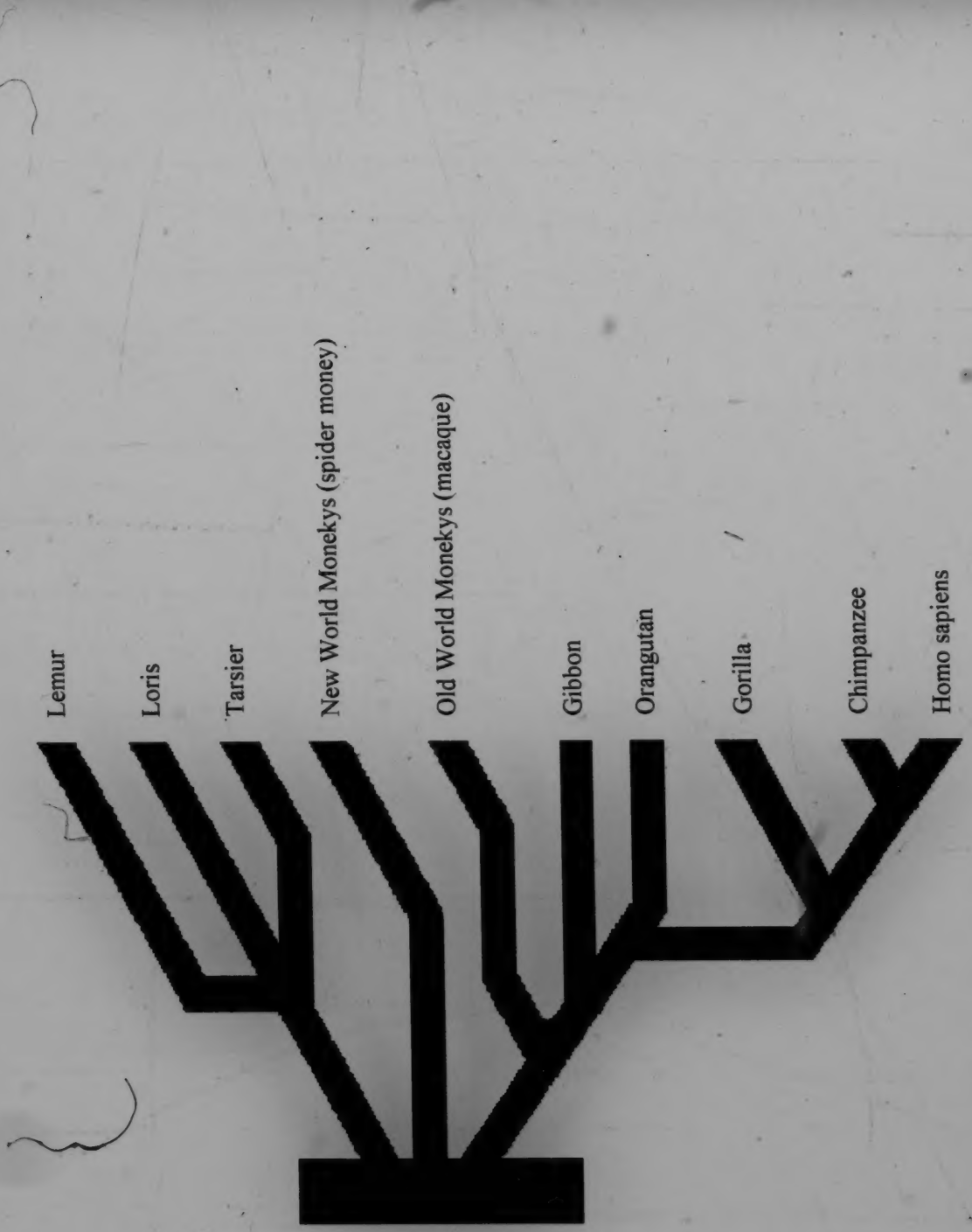


Figure 5-2: The primate lineage
(www.emory.edu/LIVING_LINKS/r/taxonomy.html).

to centromeric satellites on both chromosomal arms. In addition, FISH analyses of higher primates demonstrated that the blocks of duplicated sequences of the chromosome 10 centromere and pericentromere have been extensively rearranged over several million years. The origin of the duplication, which was traced by FISH analysis, was determined to have occurred before the divergence of orangutan (indicated by the presence of more than one signal in orangutan and only one signal in macaque). However, Jackson *et al.* also demonstrated that after the original duplication occurred, unprecedented levels (as compared to chromosomal arms) of inversion, duplication, transposition, and sequence divergence manifested. This data demonstrates the rapidly evolving nature of pericentromeric DNA.

In contrast to results reported by Regnier *et al.*, in which the duplications occurred after the divergence of orangutan from the Old World Monkeys but before the divergence from the Great Apes, Eichler *et al.* (1996) identified the duplication event that occurred between chromosome 16p11.1 and Xq28 to be more recent. Through FISH analysis Eichler *et al.* determined that the origin of the duplication was the Xq28 locus. The duplication was present in chimpanzee and gorilla, but not in orangutan and macaque. These results were expected based on the high degree of nucleotide similarity (94.6%) within the duplicated regions. The similarity between the sequences suggested to Eichler *et al.* that the duplication occurred quite recently, likely seven to ten million years ago. FISH results confirmed this prediction. In comparison, *CECR7* exons are less similar (84%-86%) to their proposed origins than the 16p11.1 duplication is to its ancestral locus, suggesting a more evolutionarily distant duplication event. Initial PCR results obtained for the intron 3 boundary were similar to those results obtained for the intron 1 boundary, and suggested that the duplication had occurred before the divergence of orangutan from the Great Ape

lineage. However, sequencing of the products revealed over the combined total of 910bp (over the intron 1 and 3 boundaries) a divergence of 0.44% between human and chimpanzee, 1.32% between human and gorilla, and 2.3% between human and orangutan, similar to recently estimated divergence rates of 1.24%, 1.62% and 3.08% respectively (Chen and Li, 2001). Based on these divergence rates, *CECR7* is predicted to have arisen from much older duplications since exons 2, 3, and 4'' are approximately 84-86% similar to their origins. Therefore, additional primer sets were used to test for the intron boundaries in macaque. For the intron 3 boundary, an additional three sets of primers were utilized for amplification of human and macaque DNA. Two of these primer sets gave sequence-verified bands in macaque. Therefore, at least the second and third *CECR7* duplicons were formed as a unit before the separation of macaque. The failure of two of the four intron 3 primer sets to produce a band in macaque is probably due to mismatches disrupting the annealing of human primers, since macaque is estimated to show approximately 5.5% sequence divergence from human (Bailey *et al.*, 2002). Mismatches may also account for the lack of macaque amplification at the intron 1 boundary, for which three different primer sets failed to give a PCR product. The sequence divergence between the human *CECR7* duplicons and their putative human origins supports the formation of *CECR7* before the divergence of Old World Monkeys (macaque) and the Great Apes. The sequence identity between the *CECR7* exon 2+3 duplicon and chromosome 16 is approximately 84.8%, while identity between the *CECR7* exon 4 duplicon and chromosome 12 is 84.5-85.8%. Thus sequence divergence in both regions is approximately 15%, much higher than the approximately 5.5% divergence expected comparing human to macaque. This high level of divergence of the *CECR7* duplicons from their ancestral loci also matches the gradient of divergence found in the

chromosome 22 pericentromeric region (Bailey *et al.*, 2001). The least diverged duplicons (< 5%) were found closest to the centromere and the most diverged were farthest from the centromere. *CECR7* is located less than 30 kb proximal to *IL17R*, the first gene outside the pericentromeric region of 22q11 (Footz *et al.*, 2001) and, therefore, *CECR7* is distant from the centromere. Sequences flanking *CECR7* also show much higher levels of divergence (approximately 15%) than typically found in pericentromeric regions. *CECR7* maps between cosmid clones 104f9 and 66F9, both of which were found to map via FISH to the equivalent of chromosome 22 in macaque (Bailey *et al.*, 2002). Therefore, *CECR7* probably formed considerably before the separation of macaque, but within the primate lineage. In contrast to *CECR7*, *NF2* is much closer to the centromere and is also much less diverged.

Although this area of research is fairly recent, most data thus far demonstrates that pericentromeric regions are rapidly evolving and that this dynamic process began fairly recently (15-33 million years ago). Obviously, some variation has been noted in the timing of the origin of specific duplicons; however, prior to the study of *CECR7*, all duplicons studied had been shown to be present in only one copy in macaque. Therefore, *CECR7* is the oldest example of a pericentromeric duplication event. The *CECR7* intron 3 boundary has been shown to have arisen from duplication events that occurred before the divergence of macaque. Although PCR results for the intron 1 boundary cannot confirm its presence in macaque, based on the degree of divergence between the origin and the duplicon, the intron 1 boundary likely formed before the divergence of macaque as well. To confirm the estimated age of the intron 1 duplication boundary, FISH could be performed on macaque metaphase spreads.

Pericentromeres are obviously junkyards for genic fragments, but may also be the birthplace of genes with novel functions. In order to determine whether *CECR7*

may function at some level in the cell, tissue-specific and species-specific expression patterns were examined by RT-PCR.

5.3 Expression Analysis of *CECR7*

Expression studies of *CECR7* indicated that this putative gene is expressed in some, but not all, tissues. Although RT-PCR products were consistently obtained from human kidney and fibroblast RNA, Northern blot analysis did not identify transcripts in any tissues tested, even in those for which expression was ascertained by RT-PCR. Northern blot analysis for *CECR7* was problematic for two reasons. Since *CECR7* is a chimeric transcript, cross-hybridization to the origins was expected. However, this was not the case, as transcripts were not identified in any tissue tested, even using a variety of different probes. Northern blot analysis is useful for expression studies in genes that are highly expressed; *CECR7* does not appear to be strongly expressed, based on Northern results and the technical difficulty experienced in examining *CECR7* expression. The integrity of the Northern blots utilized was confirmed using a control probe (GAPD).

Interestingly, *CECR7* exons 2 and 3 are derived from the *BUCS1* gene, which shows similarity to the rat kidney-specific gene. RT-PCR results verified *BUCS1* expression in kidney, as well as *CECR7* expression in kidney. Although it seems most plausible that the promoter of a gene determines its expression pattern (and the promoter of *BUCS1* was not included in the chromosome 16 duplicon), perhaps there are some functional elements within the genomic duplication from chromosome 16 that play a role in expression of the gene at the RNA level. For example, *GATA-3* is a T cell-specific transcription factor that contains a regulatory region within the first intron of the gene (Hwang *et al.* 2002). This regulatory region acts as a both a

positive and negative cis-acting element and is essential for promoter activity. The fetal liver zinc finger protein (FLIZ1) binds to the negative cis-acting element and suppresses transcription in a developmental stage-specific manner. Perhaps *CECR7* contains some sort of functional element that regulates transcription in a similar fashion. Alternatively, the duplication responsible for *CECR7* exon 1 (and the upstream genomic sequence) may also have originated in a gene that is expressed in kidney, accounting for its expression pattern, since the duplicon that carried exon 1 may also have included a promoter. There may also simply be a fortuitous promoter driving *CECR7* expression. It would be interesting to determine the expression patterns of all of the ancestral *CECR7* genes to further elucidate a mechanism by which tissue-specific expression originates (assuming that the negative RT-PCR results do indicate the absence of *CECR7* expression in some tissues). It would also be of interest to fuse the region upstream of *CECR7* to a gene with a known function and expression pattern to determine the role the region plays in gene expression. GFP or LacZ would be useful in such an experiment. The ability to determine the origin of the components of a gene is crucial in assessing how new genes are formed and how genes attain functional elements.

The limited tissue-specific expression of *CECR7* made the task of determining its expression pattern in primates more difficult. Since expression was not detected in human lymphoblasts, and only lymphoblast cell lines were available for chimpanzee, the presence or absence of transcripts in chimpanzee could not be ascertained. However, since chimpanzees are more closely related to humans than are gorillas, and since both humans and gorillas express *CECR7*, it is likely that *CECR7* expression exists in chimpanzee, as well.

There are several possibilities as to why *CECR7* expression was not detected in orangutan and macaque. Perhaps *CECR7* really is not expressed in these species. Only the duplication boundaries were cloned and sequenced, and in macaque the existence of the intron 1 boundary has still not been confirmed. Certainly *CECR7* could not be expressed in macaque if the transcription unit had not yet fully formed. It would be interesting to clone the whole gene, including the promoter, to determine structural distinctions and sequence differences, which may explain the lack of expression. A comparison of the gene structures in human, chimpanzee, gorilla, orangutan, and macaque may reveal a possible mechanism by which transcription was silenced or gained. As this region of the genome evolved, maybe a promoter was created and expression of the gene was then initiated in gorilla. This, then, leads to the question of what kind of functional element(s) gorilla may have acquired, via duplication or rearrangement events (that were shown by Jackson *et al.* (1999) to occur prolifically), that account for the expression of *CECR7* in gorilla, but not in orangutan. Alternatively, a certain sequence element may be present in the orangutan sequence that is not present in the other primates. Perhaps this sequence promotes RNA instability. Just as likely as gorilla having gained expression, orangutan may have lost *CECR7* expression over time. If *CECR7* is non-functional, expression may have been lost, foreshadowing the possible fate of the human or gorilla copy.

Moreover, there are also technical issues that must be taken into consideration in explaining these results. Quite often *in vitro* experiments do not mimic the true molecular events that occur *in vivo*. In addition, and most simply, sequence divergences in the primers, which were designed based on the human sequence, may have prevented amplification of the orangutan and macaque transcripts. A single nucleotide change at the 3' end of the primer may abolish the ability of the primer to

bind. Again, a negative PCR result cannot be unequivocally ascertained. To accurately determine whether orangutan and macaque express *CECR7*, the entire gene (both introns and exons) and the upstream sequence would need to be cloned and sequenced. This would require cloning and sequencing portions of the gene and then using species-specific primers to further clone the remaining portion of the gene. The best method to examine *CECR7* in all of the primate species may be to generate and screen primate genomic libraries. Genomic clones containing *CECR7* could then be sequenced; this approach would avoid the problem of false negatives encountered by using PCR. In addition, different primers spanning various regions of the human gene could be designed for use in RT-PCR. However, due to the apparent low level of expression, *CECR7* is not extremely amenable to RT-PCR and most regions of the gene could not be successfully amplified using human kidney RNA. Despite this difficulty, examining the expression pattern of this chimeric transcription unit is important in determining whether *CECR7* is functional.

5.4 *CECR7*: A Putative Gene

CECR7 has four small open reading frames (ORF-A is 76 amino acids, ORF-B is 103 amino acids, ORF-C is 61 amino acids, and ORF-D is 77 amino acids), two of which show similarity to known proteins (see Results pg. 43). Whether any of these ORFs are translated remains unknown. Except for ORF-C, none of the open reading frames are predicted to possess any specific motifs. *BUCS1* contains an AMP-binding domain and ORF-C overlaps with a portion of the putative AMP-binding domain encoded in *BUCS1*. However, the AMP-binding domain in ORF-C is only 40 amino acids, whereas the domain in *BUCS1* is approximately 400 amino acids. It is unlikely that the small portion of the AMP-binding domain in ORF-C is

functional. Although the maximum putative ORF encoded by *CECR7* is only 103 amino acids, some known (functional) proteins are this small. For example, histone H4 is only 12 000 kDa in size (103 amino acids), yet plays a role in the acetylation of chromatin. In addition, some mitochondrial proteins are quite small, as are ubiquitin-like proteins.

The RT-PCR product that was sequenced from gorilla has several sequence changes, but these changes do not encode a longer protein. However, since the RT-PCR product only spanned exons 1 and 2, there are possible sequence changes further downstream that may result in a longer ORF. In order to determine whether a longer open reading frame exists in gorilla (and chimpanzee), the entire gene would need to be cloned and sequenced in both species, and then the ORFs could be compared between the primates and humans. The presence of the transcript in gorilla leads to the question of whether *CECR7* is functional. Perhaps *CECR7* does function at some level and thus expression has been maintained in this species. Alternatively, the transcript may have absolutely no effect and since its presence is not harmful, aberrant transcription was neither selected for nor against.

CECR7 exons 2 and 3 are derived from the *BUCS1* gene. This gene may function in fatty acid biosynthesis or in the citric acid cycle; however, its function has yet to be tested. Perhaps ORF-C functions in a similar manner to that of *BUCS1*. However, since ORF-C encodes a protein that is only one-tenth of the size and contains only a portion of the AMP-binding domain, this seems unlikely. ORF-D spans a portion of exon 4'', which was derived from the putative gene from chromosome 12. Unfortunately, at this point the chromosome 12 gene has not been cloned any further; therefore, whether the gene is real and what its function may be remains unknown. It is also possible that this chimeric transcription unit produces a

protein with a completely novel function, or, on the contrary, is a failed experiment in evolution and is completely non-functional. An additional possibility is that *CECR7* is a functional RNA. Although it is difficult to prove a specific RNA is functional, antibodies could be produced to various regions of *CECR7* to determine if it encodes a protein. Careful design of antibodies would be required, however, to differentiate *CECR7* gene products from paralogous genes. Antibodies would be useful in determining whether *CECR7* actually produces a protein and could also be used to determine if *CECR7* interacts with other proteins.

5.5 *CECR8*: Significance of Proposed Origin

A second putative gene identified in the pericentromeric region of chromosome 22q11 was *CECR8*. Unlike *CECR7*, *CECR8* is not chimeric. *CECR8* may be an unprocessed pseudogene since the cloned portion of *CECR8* does not encode an ORF. The original location of *CECR8* was likely on the chromosome Y clone (AC006040) that contains the *CECR8*-like sequence that may encode a putative gene (see Figure 4-18). The putative Y chromosome gene from which *CECR8* was derived encodes a 130 amino acid open reading frame that is 42% similar to the chick transcription factor SOX-9 (P48434). The duplication to chromosome 22 was a genomic duplication event. This area of chromosome Y (Yp11.3) seems to be highly susceptible to duplication, as several other Y clones contain similar sequence tracts. As well, portions of *CECR8* are similar to sequence on chromosome 3 and chromosome 8. The known sequence of *CECR8* does not encode even a small ORF, suggesting that although it is transcribed, it is likely non-functional.

5.6 Expression Analysis of *CECR8*

Northern blot results indicate *CECR8* is expressed solely in testis. In addition, the two ESTs that were sequenced in order to identify part of the structure of *CECR8* were both derived from testis cDNA libraries. However, *CECR8* RT-PCR results identified four transcripts that varied in size, but only one transcript was evident on a Northern. Perhaps the alternative transcripts are not highly expressed and were therefore not visible on a Northern blot. The largest transcript cloned by RT-PCR is approximately 1.4 kb. The transcript identified by Northern blot analysis is approximately 1.6 kb. Given that *CECR8* is 89% similar to Y chromosome sequence, it is possible that the transcript identified on the Northern was actually from the Y chromosome. Perhaps the gene on the Y chromosome is not alternatively spliced, explaining the single transcript observed. Particular nucleotide changes may have resulted in the formation of novel splice acceptor and splice donor sites on chromosome 22, and led to alternative splicing in *CECR8*. The alternative transcripts (of either *CECR8* or the paralogous gene) may also be present at too low levels to be detected on by Northern blot analysis. If the band on the Northern is in fact *CECR8*, approximately 200 base pairs of this transcription unit still need to be cloned in order to fully clone the entire transcript that is expressed at the highest levels (since it is the only one that can be detected at the level of a Northern blot).

5.7 Significance of Research

The mechanism of human genome evolution is a mystery that is only beginning to come to light with the availability of extensive genomic sequence of humans and other organisms. A great deal can be learned about genome evolution by studying the evolution of genes. However, the origin of most genes is masked by evolutionary noise, which obscures the identity of the ancestral copy of a gene. It is well established (Ohno 1970) that various mechanisms of gene duplication, including whole genome duplication, are responsible for the number of genes in the human genome. However, since the last whole genome duplication was believed to have occurred over 400 million years ago (Ohno 1970), numerous nucleotide changes have obscured the origin of the genes that were created by this duplication event. However, it is possible to reconstruct the history of genes (or transcription units) such as *CECR7* that still retain significant features of their early stages. Without an understanding of where a gene originated, it is impossible to understand how particular sequences of DNA become promoters, or become functional, or how potential genes recruit functional elements.

Until recently, the extent of the evolution of the human genome was unknown. The extensive duplication and rearrangement events that Jackson *et al.* (1999) demonstrated occurring in pericentromeric regions suggests a mechanism by which new genes may be formed. Pericentromeres were previously believed to be junkyards for DNA fragments; they were known to contain gene fragments and tracts of duplicated genomic sequence, but were not generally believed to be functional. However, the discovery of *CECR7* suggests that this may not always be the case. Although it is still unclear whether *CECR7* produces a functional protein, the study of this putative gene will enhance our knowledge of genome evolution. If *CECR7* is a

functional gene, it would demonstrate the potential for the complex patchwork of duplicons at pericentromeres to act as birthplaces for novel genes limited to the Great Ape lineage. Even if *CECR7* is not presently functional, perhaps over time sequence changes will result in the formation of a new gene with a novel function. Perhaps the creation of transcription units such as *CECR7* are responsible for the phenotypic differences noted amongst the higher primates. The differences between humans and chimpanzees are so obvious, yet the degree of variation in our genomes is extremely small. Bailey *et al.* (2002) studied the entire length of chromosome 22 and identified 11 transcripts that have arisen via intrachromosomal or interchromosomal duplication. Based on the number of transcripts they identified, Bailey *et al.* estimate that since these 11 transcripts represent approximately 1% of the euchromatic genome, that approximately 1100 transcripts may have been created from duplicated sequence in the last 35 million years (the time at which macaque diverged from the Old World Monkeys).

Other examples of chimeric transcription units arising via duplication include the *jingwei* gene (Long *et al.* 1999), which was created by the integration of an *Adh* retrosequence into another gene in two *Drosophila* species (*D. teissieri* and *D. yakuba*), *ASMTL*, which formed in the pseudoautosomal region of the human sex chromosome (Ried *et al.* 1998), and a chimeric gene formed between the genes *SP100* and *HMG1L3* (Rogalla *et al.* 2000). For none of these fusion transcripts has a new function been ascribed. However, a pericentromeric duplication of an entire gene followed by evolution of a different expression pattern is illustrated by a creatine transporter gene in the 16p pericentromeric region that shows testis-specific expression, rather than the more ubiquitous expression of the ancestral gene at Xq28 (Eichler *et al.* 1997, Iyer *et al.* 1996).

Further analysis of *CECR7* may involve characterization of the ancestral copy of each exon, an examination of the possible function of this novel transcription unit, and an examination of its possible role in CES. It will be interesting to determine whether the origins of all *CECR7* exons were in expressed genes, or if random pieces of DNA integrated into the pericentromere of chromosome 22q11, acquired regulatory elements, and were subsequently expressed. The study of the ancestral genes may turn out to be significant in the field of disease genetics. For example, additional copies of a gene may cause problems in diagnosis, since separating mutations in the original gene from variations in the copies can be difficult, especially if a copy of the gene has not yet been discovered. Alternatively, the same mechanism by which pericentromeric duplications occur may be responsible (in part) for chromosomal rearrangements involved in some genetic disorders. Low copy repeats (LCRs) have been shown to be responsible for unequal cross-over during meiosis, leading to deletions and duplications of the regions contained between the LCRs. DiGeorge syndrome is caused by a deletion of part of chromosome 22q11, which results from unequal cross-over between LCRs in the region. Perhaps other repeat elements, that act as signals for the integration of duplicated fragments of DNA, may cause the same problem during meiosis. Although at this point this is all speculation, elucidation of the location of the ancestral genes and the surrounding genomic sequence may prove important in determining the etiology of various genetic disorders.

One of the great mysteries of this field is why some regions of the genome are prone to duplication and transposition. One of the big questions revolves around why pericentromeres seem to be susceptible to rearrangement events and why duplicons specifically integrate into this region of the genome, as opposed to chromosomal

arms. Clearly there must be some sort of "homing" signal to attract duplicated fragments of DNA. Possibly the same signal sequences are involved in both the duplication and integration of duplicons; perhaps the signals are different for each step of this process. Although Eichler (1998) has identified one sequence motif (CAGGG) that is involved in pericentromeric duplication and integration, there are most likely other sequences that function in the same manner.

More importantly, *CECR7* was the first reported example of a chimeric transcript formed over multiple pericentromeric duplicons (Footz *et al.*, 2001), and is also the oldest pericentromeric duplication reported to date (Bridgland *et al.* in press), since no other pericentromeric duplications have been shown to have occurred before the divergence of macaque. This transcription unit has arisen via a primate-specific duplication mechanism and allows for the study of genome evolution. Clearly the study of the evolution of pericentromeres brings about more questions than answers. However, the research involving genome evolution is rapidly increasing and hopefully new mechanisms will soon be discovered to explain the dynamic nature of the pericentromeric regions of primate chromosomes. Future work (characterization of the ancestral loci and production of antibodies) involving *CECR7* may assist in such endeavours.

Chapter 6: Introduction to *CECR1*

6.1 *CECR1*

CECR1 was identified in the McDermid lab (Riazi *et al.*, 2000) during exon trapping experiments and subsequent analysis of genomic sequence. This gene shows similarity to the Insect-Derived Growth Factor (*IDGF*), which was isolated from the flesh fly *Sarcophaga peregrina*, as well as to several other putative growth factors in various invertebrate species. For example, *Glossina morsitans morsitans* (Tsetse fly) has two genes that show similarity to *CECR1*, Tsetse salivary growth factors 1 and 2 (*TSGF-1* and *TSGF-2*) (Li and Aksoy 2000). Mollusk-derived growth factor was cloned from *Aplysia*, and is postulated to play a role during embryonic development and in CNS injury repair. In addition, six different *Drosophila* homologues have been characterized (Maier *et al.*, 2001) and all show significant similarity to the human *CECR1* gene across most of the coding region. A leader sequence has been identified at the N-terminus of the predicted CECR1 protein, suggesting that the protein is secreted. Most, but not all, of the known orthologues also have a leader sequence. The last four-fifths of the predicted protein show 26% identity and 44% similarity to the human adenosine deaminase gene (ADA). This ADA-like domain (Figure 6-1) is present in all of the invertebrate orthologues, as well. *CECR1* is hypothesized to act as a growth factor through the regulation of extracellular adenosine levels. Adenosine deaminase catalyses the deamination of adenosine to inosine. A duplication of *CECR1*, resulting in overexpression, could lead to decreased levels of extracellular adenosine, which may affect embryonic growth in individuals carrying the CESCR in greater than two copies. Zurovec *et al.* (2002) demonstrated that the addition of adenosine to imaginal disk cells inhibits growth, suggesting that the *Drosophila* orthologues of *CECR1* may promote growth by depleting extracellular adenosine.

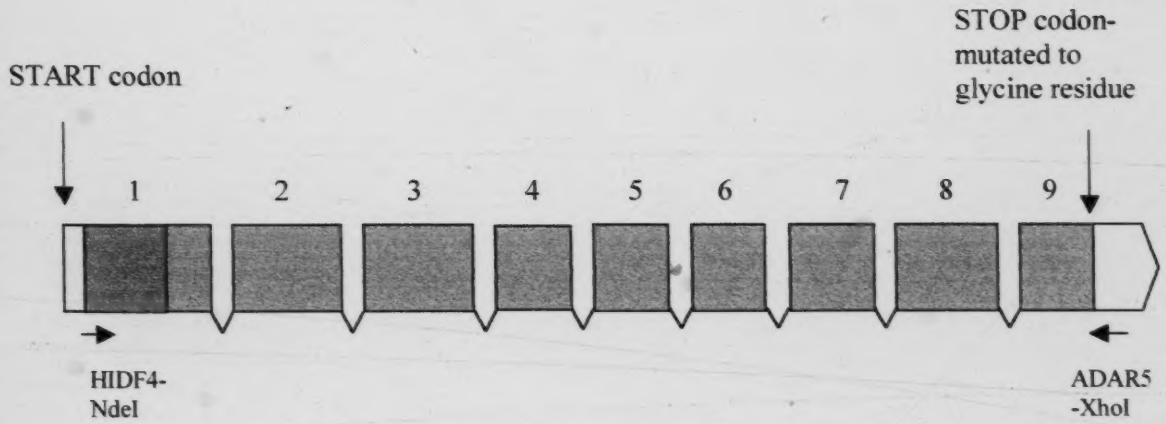


Figure 6-1: Schematic illustration of the coding region of *CECR1*. The primer set utilized in engineering the four *CECR1* constructs is labelled. The exons are colour coded to identify the three domains of the gene. The unique region is only similar to other family members in various species. The ADA-like region shows a high degree of similarity to the human adenosine deaminase gene, as well as to the 3' end of the homologous genes in other species. The unique region is denoted by ▒, the ADA-like region by ▒, and the leader sequence by □. Primer names are listed beneath the corresponding arrow. The introns are not drawn to scale.

The putative role *CECR1* may play in growth regulation and its expression profile suggests that over-expression of this gene may cause at least some of the features associated with cat eye syndrome. Northern blot analysis using the whole coding region of *CECR1* showed expression in various fetal and adult tissues (Riazi *et al.*, 2000). Strong expression was observed in adult spleen, heart, lung, lymphoblasts, and placenta, and in fetal liver, lung, and kidney. A lower level of expression was also noted in adult liver and pancreas, as well as faint expression in fetal and adult brain. RNA *in situ* hybridization results (Riazi *et al.*, 2000) revealed *CECR1* expression in the placenta, notochord, VII/VIII cranial nerve ganglion, the outflow tract of the heart, and the atrium of a 35 day old human embryo. The expression pattern of *CECR1* is consistent with some of the features associated with CES. For example, heart defects of the outflow tract are common in CES patients. Since *CECR1* is expressed in the heart during embryonic development, over-expression of *CECR1* may lead to heart defects. Likewise, expression was also noted in fetal brain. Some CES patients present with mental retardation. For the above reasons we feel that *CECR1* is a good candidate for involvement in the etiology of cat eye syndrome.

6.2 Adenosine Deaminase (ADA)

Purine nucleotides play an important role in disease etiology and in cellular homeostasis (Xu and Kellems 2000), functioning as carriers in metabolic pathways, currencies of energy, basic building blocks of RNA and DNA molecules, and as components of intracellular and intercellular signalling pathways. Purines are also involved in apoptosis. Adenine nucleosides and nucleotides (ATP and adenosine, for example) act as neurotransmitters when released from neurons into the extracellular space and have also been shown to control morphogenesis and differentiation of nerve cells during development (Michel *et al.*, 1999). These molecules prevent

degeneration of neurons in adults and may even participate in repair of brain tissue following severe ischemic brain damage.

Levels of adenosine and other purines must be tightly regulated. Following deamination of the substrate molecule, inosine and deoxyinosine are subsequently degraded to uric acid, or alternatively, are returned to the pool of purine nucleotides for further use by the cell. ADA is present in all tissues, but the levels vary widely (spanning a range greater than a 10,000 fold difference in mice) depending on cell type. Unlike *CECR1*, ADA does not have a leader sequence; therefore, ADA acts intracellularly.

Adenosine acts as an extracellular signal that elicits a variety of physiological responses by engaging cell surface receptors (Blackburn *et al.*, 1995), although not a great deal is known about the role that adenosine plays during mammalian development. However, it is apparent that ADA is important during murine fetal development (Blackburn *et al.*, 1995). ADA is expressed at very high levels in the placenta during murine development. Fetuses deficient for ADA die late in fetal development as a result of liver problems and disturbances in purine metabolism. Blackburn *et al.* (1995) genetically restored ADA to the placentas of ADA-deficient mice and were successful in correcting the majority of the problems associated with the disruption of purine metabolism. Serious liver damage was avoided and mice were rescued from perinatal lethality. Winston *et al.* (1992) showed that pharmacological inhibition of the ADA enzyme during days 7 or 8 of gestation results in malformation of almost all implanted embryos, and resorption of many of these embryos.

Severe combined immunodeficiency (SCID) in humans results from a deficiency of ADA in T-cells. In addition to a compromised immune system,

approximately 50% of SCID patients have short growth plates and some hypertrophic and necrotic chondrocytes (Aldrich *et al.*, 2000), as well as liver problems and an insufficient pulmonary system. The problems associated with SCID may be due to the accumulation of intracellular adenosine, since one of the effects of intracellular adenosine is the induction of apoptosis. It appears, then, that abnormal levels of ADA lead to abnormal growth.

In the past the ADA family has been grouped into two categories: cellular enzymes involved in purine metabolism (ADA), which are important in maintaining a competent immune system, and double-stranded RNA-specific ADAs that contain unique RNA-binding domains and are postulated to be involved in RNA editing (Keegan *et al.* 2000). The recent discovery of the invertebrate and vertebrate family of *CECR1*-like growth factor proteins creates the possibility of the existence of a third family of ADA proteins. *CECR1* and the invertebrate homologues (*IDGF* etc.) are similar to ADA in its entirety, but also contain a region upstream (of approximately 100 amino acids in *CECR1*) of the ADA-like domain that is not similar to other known proteins, but is conserved throughout this family of putative growth factors. IDGF has been shown to have ADA activity (Homma *et al.*, 2001), as has the homologous gene in *Lutzomyia longipalpis* (Charlab *et al.*, 2001). Recently, Zurovec *et al.* (2002) demonstrated that two of the *Drosophila* homologues (*ADGF-A* and *ADGF-D*) also have the ability to catalyze the deamination of adenosine to inosine *in vitro*. These genes, in addition to *CECR1*, may be classed into a third group as secretory or transmembrane ADAs (Homma *et al.* 2001). Homma *et al.* (2001) speculate that the target of chemotherapeutic drugs used to treat adult T-cell leukemia (ATL) and hairy T-cell leukemia (HTC) may not be intracellular ADA, but an extracellular or transmembrane ADA like *CECR1* (or *IDGF*), since the

chemotherapeutic agents are not able to penetrate into leukemia cells. Determining whether *CECR1* has adenosine deaminase activity is important in elucidating the mechanism by which the etiology of CES results, and may also provide important information as to how certain cancers arise. At the very least, the study of *CECR1* will further our knowledge and understanding of embryonic development.

6.3 Research Objectives

The second project involved the following goals:

1) Protein Expression of Human *CECR1* and Adenosine Deaminase:

Generate a full-length *CECR1* construct with a C-terminal HIS tag for expression in insect cells (the C-terminal HIS tag can be used for confirmation of protein expression by Western blot analysis and/or protein purification). In addition, clone the human adenosine deaminase gene (with and without a C-terminal HIS tag) into the expression vector and express the constructs in insect cells. These constructs will act as a positive control. The three constructs created can used by other members of the lab in functional assays and in the production of antibodies.

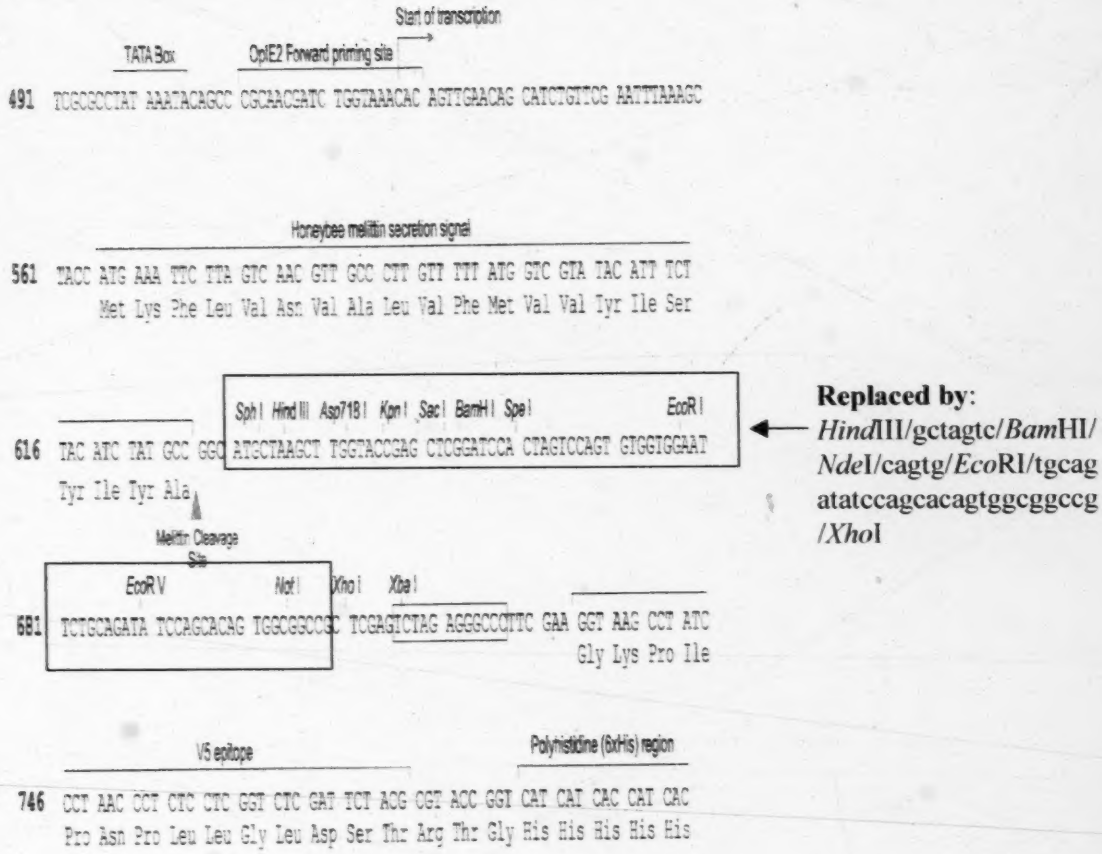
2) Test Various Constructs for ADA Activity

Test *CECR1* constructs expressed in insect cells for adenosine deaminase activity.

Chapter 7: *CECR1* Materials and Methods

7.1 Creation of *CECR1* Constructs

PCR was performed as previously described in Chapter 3 (pg 29). The *CECR1* construct was obtained by PCR from a full-length cDNA generated by Ali Riazi, a former member of the lab. To minimize PCR errors, Expand™ High Fidelity *Taq* polymerase (Roche Diagnostics) was required. Forward primers were engineered with an *Nde*I restriction site at the 5' end and reverse primers with an *Xho*I site at the 5' end of the primer to facilitate cloning into the expression vector pMIB/V5-HIS A-modified. Following PCR, products were agarose gel purified as previously described (Chapter 3) and blunt-end cloned into the pBluescript™ vector. Ligation reactions were performed according to standard protocol, except the ligation reaction occurred at room temperature for two hours (as opposed to overnight at 16°C). After PCR products were cloned into pBluescript™, they were sequence verified (as previously described) and then digested out of the vector with the appropriate enzymes. Digested products were cloned into the pMIB/V5- HIS A expression vector (Invitrogen Life Technologies), which was modified (see Figure 7-1) by Graham Banting in order to alter the reading frame so previously existing constructs could be fused in-frame with the N-terminal secretion signal and the C-terminal HIS tag and V5 epitope. PCR primers used for both cloning and sequencing can be found in Table 7-1 and 7-2 respectively. One full-length *CECR1* (CECR1-A) construct was generated, which spans the whole coding region, except for the leader sequence (Figure 6-1), as were two full-length ADA constructs, which were obtained by RT-PCR from spleen and thymus RNA. The stop codon in *CECR1-A* was mutated to a glycine residue, which allowed for the addition of a HIS tag and a V5 epitope at the C-terminal end of the protein.


Start of transcription
TATA Box
OpIE2 Forward priming site
491 TGCGCCTAT AAATACAGCC CGCAACGATC TGGTAAACAC AGTTGAACAG CATCTGTTCG AATTTAAAGC
Honeybee melittin secretion signal
561 TACC ATG AAA TTC TTA GTC AAC GTT GCC CTT GTT TTT ATG GTC GTA TAC ATT TCT
Met Lys Phe Leu Val Asn Val Ala Leu Val Phe Met Val Val Tyr Ile Ser
Sph I Hind III Asp718 I Kpn I Sac I BamH I Spe I EcoR I
616 TAC ATC TAT GCC GGC ATGCTAAGCT TGGTACCGAG CTCGGATCCA CTAGTCCAGT GTGGTGGAAT
Tyr Ile Tyr Ala
Melittin Cleavage Site
EcoR V Not I XhoI Xba I
681 TCTGCAGATA TCCAGCACAG TGGCGGCCGC TCGAGTCTAG AGGGCCCTTC GAA GGT AAG CCT ATC
Gly Lys Pro Ile
V5 epitope
Polyhistidine (6xHis) region
746 CCT AAC CCT CTC CTC GGT CTC GAT TCT ACG CGT ACC GGT CAT CAT CAC CAT CAC
Pro Asn Pro Leu Leu Gly Leu Asp Ser Thr Arg Thr Gly His His His His His

Replaced by:
*Hind*III/gctagtc/*Bam*HI/
*Nde*I/cagtg/*Eco*RI/tgcag
atatccagcacagtggcggccg
/*Xho*I

Figure7-1: A comparison of the multi-cloning sites of pMIB/V5-HIS A (Invitrogen Life Technologies) and the modified version of the vector. The modified section of the vector is labeled to the right of the original MCS. The original section of the vector, which is boxed in on the above diagram, was removed by digestion and then replaced using oligonucleotides with the new restriction sites.

The full-length ADA was cloned into pMIB/V5-HIS A-modified, as well. ADA-1 contained the stop codon, preventing addition of the 3' HIS tag. The stop codon was mutated to a glycine residue in ADA-2, allowing for the addition of the C-terminal HIS tag and the V5 epitope.

7.2 Transfection of Various Insect Cell Lines

Proteins were expressed using the INSECT SELECT™ expression system. Dr. Heather McDermid transfected three different insect cell lines with the ADA-1 construct (Sf9, Kc1, and S2 cells). Two cell lines were used for expression of ADA-2 and *CECR1-A* (S2 and Sf9). Kc1 and Sf9 cells were grown in ESF921 media, whereas S2 cells were cultured in CCM3 media. Cells were grown at 27°C in T25 flasks until the appropriate density was reached, which was 2-3 X 10^6 cells/mL for the Kc1 and S2 cell, and 4-5 X 10^6 cells/mL for the Sf9 cells. Once the appropriate cell density was reached the cells were transferred to standard six well plates. For the Sf9 cells, 4-5 X 10^6 cells were added to 1 mL of Grace's medium (Sigma) and were then incubated to allow attachment of the cells to occur. The same procedure was utilized for the S2 and Kc1 cells, except that 2-3 X 10^6 cells were added to 1 mL of Grace's medium.

During cell attachment 1.5 μg of sterile plasmid was mixed with 1 mL of Grace's medium; 10 μL of Cellfectin (Invitrogen) were then added. This mixture was incubated for 30 minutes at room temperature to facilitate the formation of DNA/liposome complexes. At the end of the 30 minute incubation, the media was removed from the attaching cells and was replaced by the mixture of DNA, Cellfectin, and Grace's media. After approximately four hours, 1.5 mL of fresh media were added to each well of the plates (ESF921 media for the SF9 and Kc1 cells and CCM3

media for the S2 cells). The plates were sealed with parafilm and the cells were grown for 48-72 hours at 27°C. After 48-72 hours the media and the cells were harvested and tested for ADA activity and transient protein expression. In addition, cells were also transferred to T25 flasks and were used in the creation of stable cell lines.

7.3 Generation of Stable Cell Lines Expressing *CECR1-A*, ADA-1, and ADA-2

Cells containing the constructs were resistant to blasticidin; therefore, blasticidin was utilized to select for recombinant cells and to generate stable recombinant cell lines expressing *CECR1-A*, ADA-1, and ADA-2. Cells from the six well plates were transferred into T25 flasks and incubated overnight at 27°C. After overnight incubation the media was removed and replaced with media containing blasticidin (ESF921 for SF9 and Kc1 cells, and CCM3 for S2 cells). The concentration of blasticidin in the media was 70 μg/mL. After the cells were grown for five days they were split using fresh media containing blasticidin. Cells were grown for two passages in media containing 70 μg/mL of blasticidin, and then maintained in media containing 10 μg/mL of blasticidin. Untransfected SF9 and S2 cells were used as controls to ensure blasticidin was in fact lethal to untransfected cells. Dr. Heather McDermid created the stable cell lines.

7.4 Isolation of *CECR1-A*, ADA-1, and ADA-2 Protein

Following transfection (48-72 hours later) cells and the media were harvested and tested for ADA activity and protein expression. Cells were removed from the six well tissue culture plates using a Pasteur pipette and transferred to 15 mL tubes. The cells and media were then centrifuged in the clinical centrifuge at speed "5" for

approximately 5 minutes. The supernatant (media) was transferred to Amicon Ultra Centrifugal Filter Device Tubes (Millipore), which were used to concentrate the supernatant down from approximately 3 mL to 0.2 mL. These columns had a molecular weight cut-off of 30 kDA, allowing for the separation of the recombinant proteins from any insect proteins that were less than 30 kDa in size. The columns containing the supernatant were centrifuged in the SH3000 rotor at 4°C for 30 minutes at 4,500 RPM. The media was tested for ADA activity immediately following concentration. The remainder of the concentrated supernatant was stored at –70°C until protein expression could be analysed by Western blot. During the concentration of the media, the cell pellets were resuspended in 0.4 mL of lysis buffer (1% IGEPAL, 50 mM Tris pH 8.0, 150 mM NaCl) and incubated on ice for 30 minutes. Cellular debris was then centrifuged out at speed "5" in the clinical centrifuge for 5 minutes. The cell lysate was also tested immediately for ADA activity.

7.5 Analysis of Protein Expression: SDS-PAGE

Protein was analyzed on SDS-PAGE gels. Protein samples were run through a 4% stacking gel and then a 7.5% separating gel (as described in the Amersham Pharmacia Biotechnology manual). 1.0 mm gels were run at 25 mA and 1.5mm gels were run at 30 mA per gel in running buffer (containing 3.02 g/L Tris HCl, 1 g/L SDS and 14.4 g/L glycine). Prior to loading, 6x SDS loading buffer (50 mM Tris-Cl (pH6.8), 100 mM Dithiothreitol, 2% SDS, 0.1% bromophenol blue, 10% glycerol) was added to the samples, which were then boiled for two minutes. The gels were either stained with silver stain (as per manufacturer's instructions for the Silver Stain Plus Kit), or electroblotted overnight (according to instructions in the Hoefer Transphor Manual) at

100 mA in Towbin buffer (containing 20% methanol, 14.42 g/L glycine, 3.05 g/L Tris-HCl and 1.0 g/L SDS) to a nitrocellulose membrane.

7.6 Western Blot Analysis

To verify transfer of proteins, the membrane was stained with Ponceau S (Boehringer-Mannheim) for 30 minutes and then destained in 5% acetic acid for 10 minutes. The membrane was then blocked with 10% skim milk powder (in 0.1% TBS-Tween) for one hour, followed by two ten minute washes in TBS-Tween. The primary antibody was diluted 1:5000 in 5% milk (the anti-V5 antibody was used for detection of the CECR1-A and ADA-2, and the anti-ADA antibody from calf-spleen (Chemicon) was used to detect expression of ADA-1) and exposed to the membrane for one hour with gentle agitation. Three 10 minute washes in TBS-Tween were performed prior to addition of the secondary antibody, which was diluted 1:5000 in 5% milk (the anti-mouse antibody for the Western blots performed with the anti-V5 antibody and the anti-rabbit antibody for the Western blot performed with the anti-ADA antibody), was added for one hour. Prior to adding the Enhanced ChemiLuminescence (ECL; Amersham Pharmacia) detection reagents, three 10 minute washes in TBS-Tween were required. 1.5 mL of each of ECL detection reagents A and B were combined and added to the blot for 90 seconds. The membrane was then sealed in a bag and exposed to X-ray film for varying amounts of time, depending on the level of protein expression.

7.7 Adenosine Deaminase Assay

Adenosine deaminase activity was measured using the LKB Biochrom Ultrospec 4050. Changes in absorbance were measured at 265 nm and a decrease in absorbance indicated the conversion of adenosine to inosine. To perform the assay, 750 μL of

100 mM potassium phosphate buffer pH 7.5, 650 μL of water, and 45 nM adenosine were added to a cuvette. Approximately 10 μL of total protein (the amount varied depending on protein concentration), which was stored on ice, was then added. The cuvette was immediately inverted to mix the reagents and the absorbance was measured at 265 nm for approximately 5 minutes (at 30 second intervals). Absorbance readings were obtained in the same manner for the adenosine deaminase (ADA) purchased from Sigma. However, because the enzyme was extremely concentrated, it was diluted prior to the assay. To dilute the enzyme, 1.2 μL of ADA was added to 1 mL of 1x PBS, and then 50 μL of the diluted enzyme was added to the reagents as earlier described. Upon failure of the ADA assay with the recombinant protein, the concentration of adenosine was altered and three different concentrations of adenosine were tested: 22.5 nM, 45 nM, and 67.5 nM.

Table 7-1: PCR Primers Used in Creating Constructs for Expression in Insect Cells

Construct	Primer Name	Primer Sequence (5'→ 3')	Tm (°C) (without restriction site)
CECR1-A	HIDF4-NdeI	GGGAATTCCATATGCGGGCGCATCTGTTGTTGAAA	66.4
	ADAR5-XhoI	CCGCTCGAGAGCTTCTCCTCCCTTTGT	62.6
ADA-1	ADA*2F1-NdeI	GGGAATTCCATATGGGGGCGCACGAGGGCACCATG	74.6
	ADA*2R1-XhoI	CCGCTCGAGACAGGGTGAAGGCTTGGAGGA	68.0
ADA-2	ADA*2F1-NdeI	GGGAATTCCATATGGGGGCGCACGAGGGCACCATG	74.6
	ADA*2R3-XhoI	CCGCTCGAGCTTGGAGGAGTGGCGTCTTCC	68.1

Table 7-2: PCR Primers Used to Sequence Constructs

Construct	Primer Name	Primer Sequence (5'→ 3')	Tm (°C)
CECR1-A	HIDR6	CCCTTTTGGCATCATCCT	62.6
	HIDF4	AGAGAAGTCAAGTGTTTA	56.9
	HIDR7	TTCCTGGTAAGTCTTCAC	60.7
	HIDF5	GCTGCTGCCGGTGTATGA	66.4
	HIDR8	GCACCTGGTTAGAGATGG	64.5
	HIDF6	AAAAGGACATCCCCATAG	60.7
ADA-1, ADA-2	ADA*2F2	TCCGCACCTGCTGGCCAACTC	71.3
	ADA*2R2	CCAGGGGCAGATCTCGAAGTG	69.7
All constructs	T7	GTAATACGACTCACTATAGGGC	64.9
	T3	AATTAACCCTCACTAAAGGGAA	61.8
	Op1E2 F	CGCAACGATCTGGTAAACAC	64.8
	OpIE2 R	GACAATACAAACTAAGATTTAGTCAG	62.8

Chapter 8: *CECR1* Results

8.1 Expression of Recombinant Proteins and Analysis of ADA Activity

The goal of this project was to determine whether *CECR1* has adenosine deaminase activity. Since all of the adenosine deaminase catalytic residues are conserved in *CECR1* (Figure 8-1), ADA activity was expected for this protein. In order to test this gene for ADA activity, the whole coding region (minus the leader sequence) was cloned into the insect expression vector pMIB/V5-HIS A (modified). CECR1-A spanned 480 of the 511 amino acids of the coding region. The first 31 amino acids of *CECR1* were eliminated from the construct because the expression vector was designed to secrete the expressed protein. In addition, the stop codon was mutated to a glycine residue to allow for the addition of a V5 epitope and a HIS tag at the C-terminus of the protein (Figure 7-1).

Since the only positive control available was purified enzyme (Sigma), a positive control that was expressed under the same conditions as the experimental constructs was required to assess the validity of results obtained from protein expression in insect cells. Human adenosine deaminase (ADA-1, ADA-2) was cloned by RT-PCR from two separate tissues, spleen and thymus, as illustrated in Figure 8-2. Subsequent sequence analysis identified a two nucleotide mutation that was observed in several clones. The first nucleotide change was the third base of the codon and did not change the amino acid. However, the second mutation resulted in a non-conservative amino acid change from a lysine to a glutamic acid residue. Since these particular mutations were found in clones obtained from independent RT-PCR reactions, and the clones were derived from two different tissues, database searches were performed and several human, rat, and cow ESTs were identified that contained the same two nucleotide changes, indicating these changes were not mutations; rather,

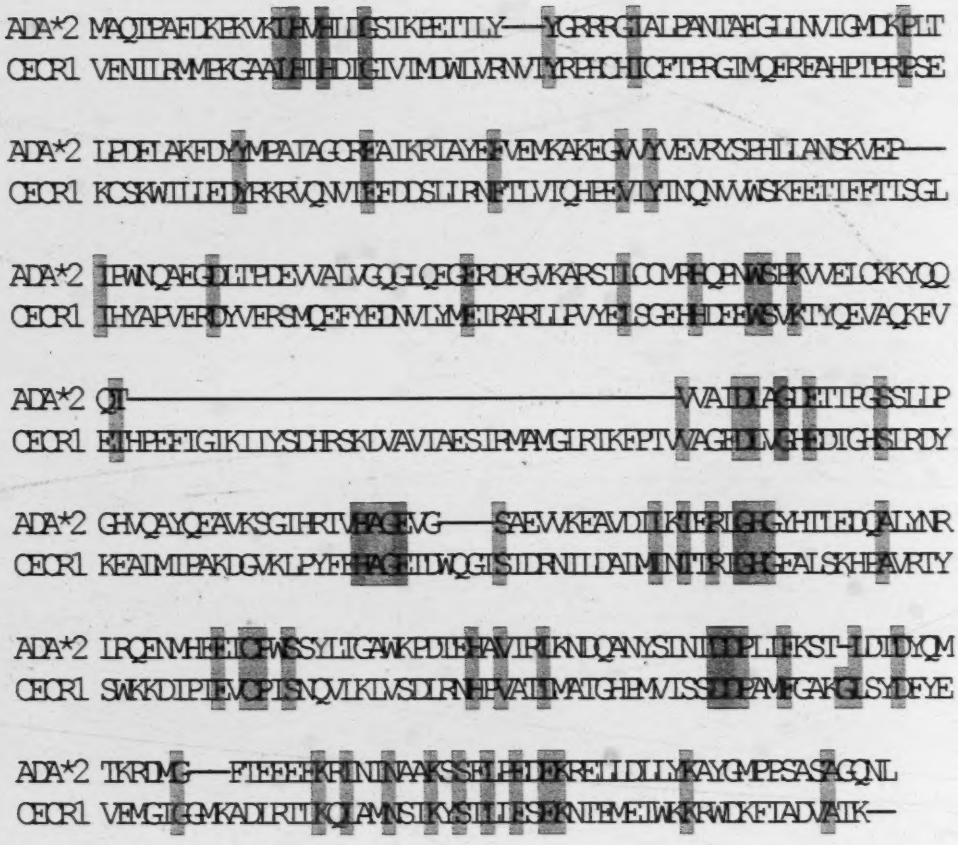


Figure 8-1: A comparison of the amino acid sequences of the human adenosine deaminase gene (ADA*2) and the ADA-like region of *CECR1*. Conserved amino acids are coloured in blue and conserved catalytic sites are shaded in red.

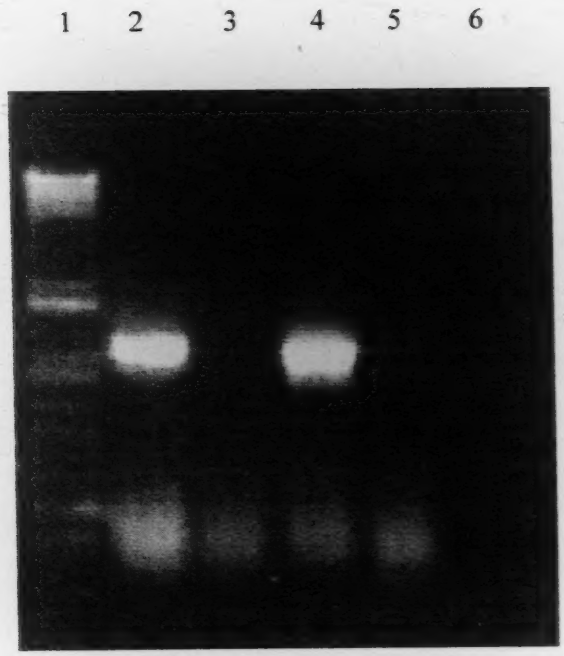


Figure 8-2: ADA-2 RT-PCR results. Labeled lanes are as follows: Lane 1: 1kb+ ladder; Lane 2: spleen; Lane 3: spleen negative control; Lane 4: thymus; Lane 5: thymus negative control; Lane 6: water control. RT-PCR negative controls contained RNA, but no reverse transcriptase, to rule out genomic DNA contamination.

they were polymorphisms (ADA*2). Both ADA constructs spanned the coding region from start to stop. ADA-1 contained the native stop codon. Since the HIS tag is added to the C-terminus of the expressed protein, ADA-1 did not contain a HIS tag. However, ADA-2 was tagged at its C-terminus with a V5 epitope and a HIS tag, as the stop codon was mutated to a glycine residue during PCR.

The media and the insect cells were harvested 48-72 hours after transfection. The media was concentrated down from approximately 3 mL to 0.2 mL, which was the minimum amount the concentrator columns would allow to be isolated. In addition, the cells from three transfections of the same construct in the same cell line were pooled to increase the amount of expressed protein obtained. After lysis of the cell pellets, the ADA assay was performed on both the concentrated media and on the cell pellet lysates. Adenosine deaminase that was purchased from Sigma was employed as a positive control to ensure that the assay could detect the conversion of adenosine to inosine. In addition, the Sigma ADA was used as a control during the lysis of the cell pellets, during the concentration of the media, and was also incubated with cultured cells for 24 hours to ensure that other proteins produced by the insect cells did not inhibit activity. Likewise, the controls served to ensure that ADA activity was not being lost over time and that the enzyme was active when cultured at 27°C. ADA obtained from Sigma consistently demonstrated activity, even after incubation at 27°C for 24 hours. In addition, lysis of the cell pellet in the NP-40 buffer did not destroy Sigma ADA activity, and activity was also preserved during media concentration (see Table 8-1). Therefore, the methods by which the recombinant protein was isolated and tested for ADA activity should theoretically maintain the integrity of the expressed protein. The minimum amount of ADA required for ADA activity to be evident in a spectrophotometric assay was also

assayed using a serial dilution of the Sigma ADA. Results indicated that only 16ng of protein was required in order for the conversion of adenosine to inosine to be detected (see Table 8-2).

Western blot results indicated that ADA-2 was highly expressed in S2 and SF9 cells (Figure 8-3). Greater protein expression was observed in the cell pellets than in the concentrated media of Sf9 cells. ADA-2 expression was not evident in the media from the S2 cells. CECR1-A expression was detected at an extremely low level in the cell pellet (data not shown). Upon creation of a stable cell line, however, many cells survived, indicating that they had been transformed and were expressing CECR1-A. Perhaps CECR1-A expression was time dependent and increased after the cells were harvested at 72 hours. Expression of ADA-1 could not be ascertained. Since ADA-1 was not tagged, a different antibody (anti-bovine ADA from Chemicon) was used to identify ADA-1 by Western blot analysis. Unfortunately, the antibody purchased from Chemicon was the purest antibody available (but is only semi-purified) and was not specific enough to identify ADA-1. Western blot results using the Chemicon antibody (at various dilutions) consistently identified a smear of bands, indicating the antibody was non-specific. Therefore, ADA-1 expression could not be ascertained.

The ADA assay was performed on both the media and cell pellet extracts derived from transient transfections (not from the stable cell lines) for all constructs in the various cell lines. Varying concentrations of adenosine were used (22.5, 45, and 67.5 nM), as were varying amounts of protein (10-25 μL, depending on the absorbance). ADA from Sigma was used to test the assay prior to assaying the

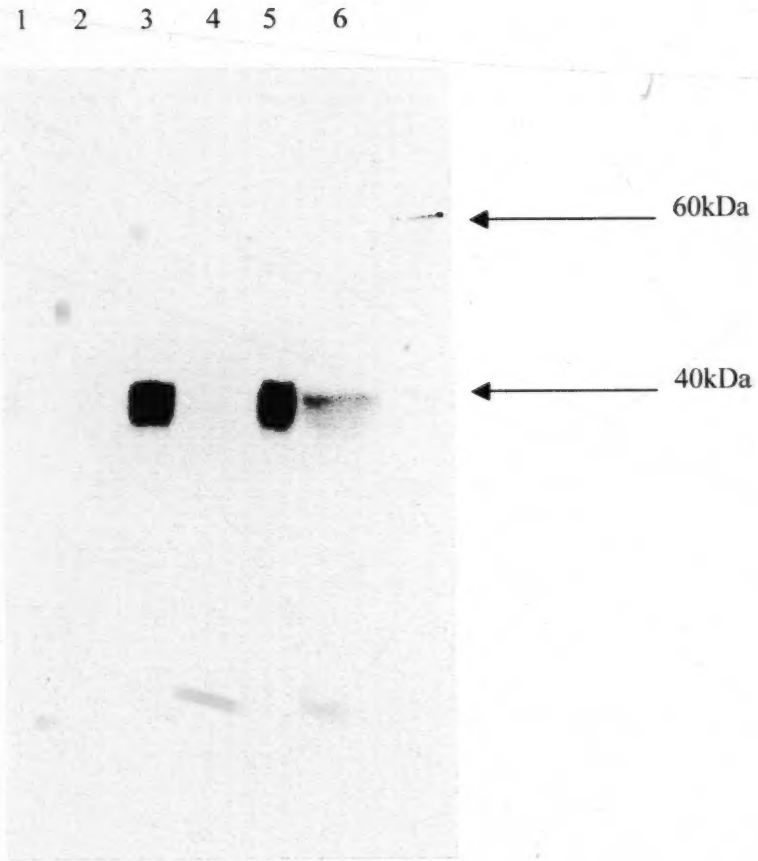


Figure 8-3: Western blot results for insect expressed proteins. Results demonstrate that ADA-2 was expressed in both cell lines, but substantial secretion into the media only occurred in the SF9 cell line. For both cell lines, protein expression (or retention) was much greater in the cell pellet. The Western blot was probed with the anti-V5 antibody. Labeled lanes are as follows: Lane 1: Kc-1 pellet; Lane 2: Kc-1 media; Lane 3: S2 pellet; Lane 4: S2 media; Lane 5:SF9 pellet; Lane 6: SF9 media. The 60kDA marker of the ladder is labeled, as well.

experimental samples. Each time the assay was performed with the Sigma ADA, results were as expected. Over a period of approximately 5 minutes the absorbance at 265 nm decreased substantially, indicating that adenosine was converted to inosine. However, ADA activity was not detected in the supernatants or the cell pellets of CECR1-A, ADA-1, or ADA-2. The assay was performed on several occasions using protein from transient expressions, and each time the results were consistent (Table 8-3). Although a slight variation was often noted in the absorbance at 265 nm, the fluctuations resulted in the absorbance increasing and decreasing by equal amounts. In addition, the negative controls, which were cells from the various lines that had not undergone transfection, demonstrated the same pattern of fluctuation.

Table 8-1: Assay Results Using Commercial ADA

Experiment	Initial Absorbance	Final Absorbance	Change in Absorbance (over 3 minutes)
Commercial ADA (in media from insect cells)	1.150	0.905	0.245
Commercial ADA in lysis buffer	0.986	0.782	0.204
Commercial ADA concentrated with supernatant	0.998	0.722	0.266
Insect media (S2) spiked with Commercial ADA (30 minute incubation at 27°C)	0.875	0.646	0.229
Insect media (S2) spiked with Commercial ADA (overnight incubation at 27°C)	0.926	0.643	0.283
Insect media (Sf9) spiked with Commercial ADA (30 minute incubation at 27°C)	1.042	0.835	0.207
Insect media (Sf9) spiked with Commercial ADA (overnight incubation at 27°C)	1.081	0.809	0.272

Table 8-2: ADA Assay Results using a Dilution Series of Commercial ADA

Protein Concentration	Initial Absorbance	Final Absorbance	Change in Absorbance (in 4 minutes)
160 ng	0.513	0.254	0.259
16ng	0.351	0.263	0.088
1.6ng	0.298	0.295	0.003

Table 8-3: ADA Assay Results for the Insect Expressed Proteins

Construct/Cell Line	**Initial Absorbance**	**Final Absorbance**	**Change in Absorbance (over 3 minutes)**
ADA-1 S2 media	0.832	0.827	0.005
ADA-1 S2 cell pellet	1.030	1.028	0.002
ADA-1 Sf9 media	0.711	0.712	+0.001
ADA-1 Sf9 cell pellet	0.949	0.956	+0.007
ADA-2 S2 media	0.790	0.788	0.002
ADA-2 S2 cell pellet	1.034	1.039	+0.005
ADA-2 Sf9 media	0.687	0.689	+0.002
ADA-2 Sf9 cell pellet	1.072	1.089	+0.017
CECR1-A S2 media	0.554	0.556	+0.002
CECR1-A S2 cell pellet	0.895	0.898	0.003
CECR1-A Sf9 media	1.033	1.043	+0.010
CECR1-A Sf9 cell pellet	1.027	1.027	0.000

Chapter 9: *CECR1* Discussion

9.1 Expected Results versus Apparent Results: *CECR1-A*

Orthologues of *CECR1* have been identified in many species, including several insect species (Li and Aksoy 2001, Maier *et al.* 2001, Charlab *et al.* 2001, Homma *et al.*, 2001). Several of the *Drosophila* homologues have demonstrated ADA activity (Zurovec *et al.* 2002), as well as the two *Glossina m. mortisans* (tsetse fly) homologues (Li and Aksoy 2001), and the homologue in *Lutzomyia longipalpis* (Charlab *et al.* 2001). Based on the high degree of similarity between these insect homologues and *CECR1*, and the similarity of the C-terminus to the human adenosine deaminase gene, *CECR1* was expected to convert adenosine to inosine and to demonstrate adenosine deaminase activity. However, when CECR1-A, ADA-1, and ADA-2 were expressed in insect cells, ADA activity was not apparent. Although substantial expression of ADA-2 was confirmed by Western blot analysis, expression of CECR1-A and ADA-1 could not be confirmed. Recombinant protein expression is expected from these two constructs, since stable cell lines were created using blasticidin, which is a selectable marker that selects for cells containing the pMIB/V5-HIS A vector. However, at this time, the viability of the spectrophotometric assay in testing for ADA activity of insect-expressed human proteins can only be based on results obtained for ADA-2, since expression has been definitively confirmed.

There are several reasons that ADA activity may not have been apparent for the INSECT SELECT™ expressed human adenosine deaminase (ADA-2) protein. Firstly, perhaps there was not enough protein for ADA activity to be evident. The minimum amount of purified protein required to observe ADA activity is 16 ng. Although there was more than 16 ng of recombinant protein in the assays (based on Western results), the amount of commercially purified protein and the quantity of

expressed protein required are likely not equivalent. The Sigma enzyme is manufactured to be stable at 4°C and its efficiency has been tested. Perhaps the glycerol solution in which it is stored adds to its stability and enhances its activity. Although mock experiments were performed using the Sigma ADA to test whether the lysis buffer would destroy adenosine deaminase activity or whether activity was lost during the concentration of the media, these results do not necessarily correlate with the expression of recombinant proteins in insect cells, as the expressed proteins are not stabilized commercially. Since the commercial enzyme is marketed as a reliable standard to test for ADA activity, it is reasonable to assume it is engineered to withstand some harsh conditions. On the other hand, the INSECT SELECT™ expressed proteins are not.

Moreover, it is possible that ADA-2 did have adenosine deaminase activity, but either the assay was not sensitive enough to detect the activity, or the conversion of adenosine to inosine was obscured by the remaining adenosine. For example, if there was only a small quantity of active protein and a great deal of substrate (of which the absorbance is being measured), then only a few substrate molecules would be converted to inosine. Therefore, the remaining adenosine would still cause the absorbance to remain stable, even though a small amount of inosine had been generated. Perhaps decreasing the concentration of adenosine will also allow for a small change in the adenosine and inosine levels to be detected. Likewise, a more sensitive assay may prove useful in elucidating whether the insect expressed human ADA is in fact active. The conversion of adenosine to inosine can be monitored using C-14 labelled adenosine by chromatography.

Perhaps the C-terminus V5 epitope/HIS tag inhibited activity. Kawamura *et al.* (1999) demonstrated that Imaginal Disk Growth Factors (*IDGF*) 1 and 2 (unrelated

to *CECR1*) did not show growth factor activity when expressed with an N-terminus HIS tag. These results indicate that the addition of only a few amino acids may alter the function of a protein. Since both CECR1-A and ADA-2 contained a 3' HIS tag (and the stop codon was mutated to a glycine residue, which may affect function, as well), this may explain why neither construct demonstrated activity. However, ADA-1 did not contain the HIS tag and did contain the native stop codon, and did not demonstrate the ability to deaminate adenosine. Again, however, ADA-1 protein expression was not confirmed and the results are, therefore, not interpretable.

In addition, insect cells may not be a viable system for functional analyses of this human protein. Human proteins may require posttranslational modifications (*CECR1* is predicted to require glycosylation), some of which are not performed within insect cells. Perhaps the recombinant proteins were not modified correctly following translation, which would result in a loss of enzymatic activity. Zurovec *et al.* (2002) obtained functional proteins (with ADA activity) by using baculovirus to express the *Drosophila* homologues of *CECR1*. Since these insect genes were expressed in insect cells, they were likely modified correctly and, therefore, were functional. Ideally, human proteins should be expressed in mammalian cells to ensure the proper posttranslational modifications occur. However, functional adenosine deaminase proteins have been obtained via bacterial protein expression. For example, Yeung *et al.* (1985) cloned the murine adenosine deaminase cDNA into a bacterial vector and assayed for complementation in ADA deficient *E. coli*. This group demonstrated functional mammalian adenosine deaminase could be produced in prokaryotes. Despite the results of this experiment, more often than not, groups studying the human copy of adenosine deaminase express the protein in mammalian cells, ensuring functional activity of the protein.

9.2 Future Directions

Efforts are currently underway to continue analysis of the INSECT SELECT™ expressed proteins. The creation of stable cell lines allows for a greater production of protein, which may circumvent the possibility of there being too little protein to detect adenosine deaminase activity. The experiment is being tweaked in other ways, as well. The concentration of adenosine will be varied more widely, and perhaps the ADA assay itself will be changed to a more sensitive detection system. In addition, the proteins with the C-terminus HIS tags may be purified, which will increase the relative concentration of the recombinant protein. Although two constructs were obtained for ADA, one with a HIS tag and one without, the only CECR1-A construct presently expressed in the insect cells has a C-terminal HIS tag. It is necessary to create a *CECR1* construct without a HIS tag in the event that the HIS tag does inhibit adenosine deaminase activity.

If functional proteins are still not obtained using the INSECT SELECT™ system, these constructs will be expressed in mammalian cells, using a constitutive promoter. I expressed the human copy of adenosine deaminase in mammalian COS cells (results not reported); however, ADA activity was not detected 24, 48, or 72 hours after transfection. Since the mammalian expressed protein was not tagged, the anti-ADA antibody (from Chemicon) was used, but due to its low level of purity, expression of this construct could not be confirmed. Several modifications could be made to ensure a functional protein is obtained from mammalian cells. Firstly, a working antibody for both ADA and *CECR1* are imperative. Secondly, various cell lines should be utilized for protein expression.

Complementation may also be an effective tool to assay for ADA activity. The human ADA and *CECR1* constructs could be transformed into ADA-deficient

bacterial cell lines. Only those cells with the ability to convert adenosine to inosine will survive. This approach was taken by Ingolia *et al.* (1986). Ingolia *et al.* used genetic complementation of ADA-deficient bacteria to isolate copies of mammalian ADA cDNA sequences by cloning various ADA cDNAs into a vector and then transforming ADA-deficient bacteria with the constructs.

Not only have several of the insect homologues demonstrated ADA activity, but they have also been shown to act as growth factors (Homma *et al.* 1996, Homma *et al.* 2001). Homma *et al.* postulate that the adenosine deaminase activity also plays a role in the ability of the protein to promote the growth of cells. They hypothesize that when adenosine binds to a membrane receptor it is converted to inosine; this conversion may then act as a signal that is transmitted into the cells, resulting in the proliferation of cells. However, Zurovec *et al.* (2002) demonstrated that the addition of adenosine to human endothelial cells results in their proliferation, whereas ADA inhibits proliferation. At higher concentrations of adenosine, however, apoptosis is induced in endothelial cells. Similarly, adenosine (in the presence of ADA inhibitors) is toxic to human carcinoma cell lines. These seemingly disparate results indicate that the effect of adenosine may be dependent on cell type. If *CECR1* does in fact demonstrate the ability to convert adenosine into inosine, the next logical step will be to test this human protein for growth factor activity. The same constructs engineered to test for adenosine deaminase activity will be useful for this task, as well. Perhaps testing for growth factor activity using a variety of cell types may allow us to circumvent the ADA assay. Adenosine could be added to cells expressing CECR1 and ADA. The growth properties of the cells could then be examined. Efforts may also focus on identifying interacting proteins, if *CECR1* does in fact interact with other proteins.

9.3 Significance of Research

CECR1 is a good candidate for involvement in cat eye syndrome. The study of this gene and the unravelling of the mystery of its role in mammalian development lies in determining its functional significance. The expression pattern of *CECR1* suggests it may play a role in the development of the heart (Riazi *et al.* 2000). Heart defects are a common feature of CES. If *CECR1* does act as a growth factor by controlling the level of extracellular adenosine, perhaps a duplication of this gene leads to depleted adenosine levels, which results in a defect in the vascularization of the vessels leading to and from the heart muscle (Leibovich *et al.* 2002). Perhaps the effect of having four copies of *CECR1* stems from a downstream interaction with the receptor of this gene. The receptor then may play a role in the manifestation of the phenotypic characteristics of CES. Regardless, the study of the functionality of *CECR1* will lead to novel interpretations regarding the role of growth factors in embryonic development and may help elucidate the biological etiology of cat eye syndrome.

Chapter 10: Conclusion

10.1 Overall Significance of Research: A Final Summary

The McDermid lab's ultimate goal is to elucidate the molecular etiology of cat eye syndrome. Results of these studies involving *CECR7*, *CECR8*, and *CECR1* are small steps along the path to realizing this particular goal. Analysis of *CECR8* indicates this putative gene is likely to be a transcribed pseudogene, ruling out the likelihood of it acting in a dose dependent manner and in the causation of CES. Expression studies indicate a very limited expression profile, identifying high levels of expression in testis (based on Northern and RT-PCR results) and low expression levels in fetal brain. Since CES patients do not normally present with testicular defects, the tissue-specific expression pattern also suggests this transcript is an unlikely candidate for involvement in cat eye syndrome.

Similarly, but not as clearly, *CECR7* is now believed to be less likely a candidate for causing features of CES (than it was at the time this project was initiated) based on the fact that it is unclear whether this putative gene is functionally significant. However, *CECR7* is expressed in the kidney (demonstrated by RT-PCR), one of the organs that is often defective in CES patients. The small ORFs cannot be deemed functionally insignificant until proven otherwise, leaving room for the possibility that the duplication of *CECR7* may be involved in causing kidney defects in CES patients. *CECR7* may function as an RNA, but proving this could be extremely difficult. Interestingly, Townes-Brocks syndrome, which was discussed in the introduction, has an overlapping phenotype with CES. The deletion of *SALL1* has been implicated in causing the features of this syndrome (Kohlase *et al.* 1998). Recently *SALL1* was shown to bind to the pericentromeric region of chromosome 22 (Netzer *et al.* 2001); the binding of this transcription factor inhibits transcription of

genes in this region. Perhaps when *SALLI* is deleted, *CECR7* transcription is no longer repressed, resulting in over-expression. Over-expression of *CECR7* may lead to the features of CES that overlap with Townes-Brocks syndrome. At this point this theory is only speculation; however, future work may include an examination of which genes are transcriptionally repressed by the binding of *SALLI* in the pericentromere of chromosome 22.

CECR1 is unlikely to be solely responsible for CES. Although *CECR1* is duplicated in all patients that have been tested to date, the CES phenotype is quite variable, suggesting a synergistic effect may occur. However, its expression profile (Riazi 2000, summarized in Section 6.1) implies it may play a role in the heart defects common to CES patients. Furthermore, adenosine is decisively involved in the developing fetal heart (reviewed in Rivkees *et al.*, 2001). Rivkees *et al.* review the results from several experiments that indicate adenosine levels must be tightly regulated during embryonic development, as studies of cultured embryos reveal that the dominant humoral regulator of cardiac function in embryos is adenosine. In the presence of adenosine re-uptake inhibitors, fetal heart rates plummet. This is in contrast to the role a duplication of *CECR1* would presumably play; increased dosage of *CECR1* would presumably lead to deceased adenosine levels, if this gene does in fact act in the same manner as the human adenosine deaminase. Similarly, increased adenosine levels in the cardiac region result in bradycardia, explaining the drop in embryonic heart rate, and can also lead to asystole, a condition that may lead to sudden fetal death. High levels of adenosine are also associated with a reduction in the size of the ventricles, intrauterine growth failure, and thin ventricular walls. Ali Riazi, at the University of Toronto, is presently testing the putative function of *CECR1* in the heart. He has placed *CECR1* under the control of both a heart specific

promoter and a promoter with ubiquitous expression in mice (Riazi and Buchwald 2001). Transgenic lines expressing *CECR1* under the control of both promoters demonstrated a high rate of embryonic and neonatal lethality. Both lines presented with cardiac hypertrophy, which resulted from an enlargement of the atria and the right ventricle. Kidney defects (hypoplastic or underdeveloped kidneys) were noted in transgenic mice expressing *CECR1* under the control of the promoter with ubiquitous expression. These results are promising and suggest that *CECR1* may be responsible for at least the common heart defects, and perhaps the kidney defects, with which CES patients present. However, since a mouse orthologue of *CECR1* has not yet been identified, the correlation of the phenotype in mice expressing *CECR1* to the phenotype observed in CES patients remains unknown.

Although much work is still required to ascertain exactly which genes are involved in the etiology of cat eye syndrome, progress is being made toward meeting the McDermid lab's ultimate goal. Perhaps other candidates may be involved in causing some of the features of CES. Perhaps CES is caused by a single gene in the cat eye syndrome critical region. Regardless, many interesting discoveries have occurred through this study and through the research of other members of the lab. The focus of the research may be to determine the role that genes in the CESCR play in CES, but these studies will continue to expand our knowledge of the development of the human fetus and add to the growing interest in genome evolution.

References

Aldrich MB, Blackburn MR, Kellems RE. (2000) The importance of adenosine deaminase for lymphocyte development and function. Biochemical and Biophysical Research Communications 272(2): 311-315.

Altschul, SF, Madden T, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402.

Arnold PD, Siegel-Bartelt J, Cytrynbaum C, Teshima I, and Schachar R (2001) Velo-Cardial-Facial Syndrome: Implications of microdeletion 22q11 for schizophrenia and mood disorders. American Journal of Medical Genetics (Neuropsychiatric Genetics) 105: 354-362.

Ausubel, F *et al.* (1991) *Current Protocols in Molecular Biology.* Wiley Interscience, New York.

Bailey JA, Yavor AM, Viggiano L, Misceo D, Horvath JE, Archidiacono N, Schwartz S, Rocchi M, Eichler EE. (2002) Human-specific duplication and mosaic transcripts: the recent paralogous structure of chromosome 22. American Journal of Human Genetics 70(1): 83-100.

Baldini A. (2002) DiGeorge syndrome: the use of model organisms to dissect complex genetics. Human Molecular Genetics 11(20): 2363-2369.

Berends MJW, Tan-Sindhunata G, Leegte B,Van Essen AJ. (2001) Phenotypic variability of cat eye syndrome. Genetic Counselling 12(1): 23-34.

Blackburn MR, Wakamiya M, Caskey CT, Kellems RE. (1995) Tissue-specific rescue suggests that placental adenosine deaminase is important for fetal development in mice. Journal of Biological Chemistry 270(41): 23891-23894.

Bridgland L, Footz TK, Kardel MD, Riazi MA, McDermid HE. Three Duplicons Form a Novel Chimeric Transcription Unit in the Pericentromeric Region of Chromosome 22q11. Human Genetics, in press.

Burge C, Karlin S. (1997) Prediction of complete gene structures in human genomic DNA. Journal of Molecular Biology 268: 78-94.

Carey AH, Kelly D, Halford S, Wadey R, Wilson D, Goodship J, Burn J, Paul T, Sharkey A, Dumanski J, et al. (1992) Molecular genetic study of the frequency of monosomy 22q11 in DiGeorge syndrome. American Journal of Human Genetics 51(5): 964-970.

Charlab R, Valenzuela JG, Andersen J, Ribeiro JM. (2001) The invertebrate growth factor/CECR1 subfamily of adenosine deaminase proteins. Gene 267(1): 13-22.

Chen FC, Li WH. (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. American Journal of Human Genetics 68(2): 444-456.

Dunham I, Shimizu N, Roe BA, Chissoe S, Hunt AR, Collins JE, Bruskiewich R, Beare DM, Clamp M, Smink LJ, *et al.* (1999) The DNA sequence of human chromosome 22. Nature 402: 489-495.

Edelmann L, , Spiteri E, Koren K, Pulijaal V, Bialer MG, Shanske A, Goldberg R, Morrow BE. (2001) AT-rich palindromes mediate the constitutional t(11;22) translocation. American Journal of Human Genetics 68(1): 1-13.

Eichler EE, Lu F, Shen Y, Antonacci R, Jurecic V, Doggett NA, Moyzis RK, Baldini A, Gibbs RA, Nelson DL. (1996) Duplication of a gene-rich cluster between 16p11.1 and Xq28: a novel pericentromeric-directed mechanism for paralogous genome evolution. Human Molecular Genetics 5(7): 899-912.

Eichler EE, Budarf ML, Rocchi M, Deaven LL, Doggett NA, Baldini A, Nelson DL, Mohrenweiser HW. (1997) Interchromosomal duplications of the adrenoleukodystrophy locus: a phenomenon of pericentromeric plasticity. Human Molecular Genetics 6(7): 991-1002.

Eichler EE. (1998) Masquerading repeats: paralogous pitfalls of the human genome. Genome Research 8(8): 758-762.

Eichler EE. (1999) Repetitive conundrums of centromere structure and function. Human Molecular Genetics 8(2): 151-155.

Eichler EE, Archidiacono N, Rocchi M. (1999) CAGGG repeats and the pericentromeric duplication of the hominoid genome. Human Molecular Genetics 9(11): 1048-1058.

Fairbrother WG, Yeh RF, Sharp PA, Burge CB. (2002) Predictive identification of exonic splicing enhancers in human genes. Science 297(5583): 1007-1013.

Footz TK, Brinkman-Mills P, Banting GS, Maier SA, Riazi MA, Bridgland L, Hu S, Birren B, Minoshima S, Shimizu N, Pan H, Nguyen T, Fang F, Fu Y, Ray L, Wu H, Shaull S, Phan S, Yao Z, Chen F, Huan A, Hu P, Wang Q, Loh P, Qi S, Roe BA, McDermid HE. (2001) Analysis of the cat eye syndrome critical region in humans and the region of conserved synteny in mice: a search for candidate genes at or near the human chromosome 22 pericentromere. Genome Research 11(6): 1053-1070.

Franklin RC, Parslow MI. (1972) The cat-eye-syndrome. Review and two further cases occurring in siblings with normal chromosomes. Acta Pediatric Scandinavia 61: 581-586.

Frohman MA, Dush MK, Martin GR. (1988) Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer. Proceedings of the National Academy of Science of the USA 85(23): 8998-9002.

Gong W, Gottlieb S, Collins J, Blescia A, Dietz H, Goldmuntz E, McDonald-McGinn DM, Zackai EH, Emanuel BS, Driscoll DA, Budarf ML. (2001) Mutation analysis of TBX1 in non-deleted patients with features of DGS/VCFS or isolated cardiovascular defects. Journal of Medical Genetics 38(12): E45.

Gromak N, Smith CW. (2002) A splicing silencer that regulates smooth muscle specific alternative splicing is active in multiple cell types. Nucleic Acids Research 30(16): 3548-3557.

Haab O (1878) Beitrage zu den angeoborenen Fahlern des Auges. Von Graefe's Arch Opthalmology 24: 257-281.

Homma K, Matsushita T, Natori S. (1996) Purification, characterization, and cDNA cloning of a novel growth factor from the conditioned medium of NIH-Sape-4, an embryonic cell line of Sarcophaga peregrina (flesh fly). Journal of Biological Chemistry 271(23): 13770-13775.

Homma KJ, Tanaka Y, Matsushita T, Yokoyama K, Matsui H, Natori S. (2001) Adenosine deaminase activity of insect-derived growth factor is essential for its growth factor activity. Journal of Biological Chemistry 276(47): 43761-43766.

Horvath JE, Viggiano L, Loftus BJ, Adams MD, Archidiacono N, Rocchi M, Eichler EE. (2000a) Molecular structure and evolution of an alpha satellite/non-alpha satellite junction at 16p11. Human Molecular Genetics 9(1): 113-123.

Horvath JE, Schwartz S, Eichler EE. (2000b) The mosaic structure of human pericentromeric DNA: a strategy for characterizing complex regions of the human genome. Human Molecular Genetics 10(6): 839-852.

Hwang ES, Choi A, Ho IC. (2002) Transcriptional regulation of GATA-3 by an intronic regulatory region and fetal liver zinc finger protein 1. Journal of Immunology 169(1): 248-253.

Iyer GS, Krahe R, Goodwin LA, Doggett NA, Siciliano MJ, Funanage VL, Proujansky R. (1996) Identification of a testis-expressed creatine transporter gene at 16p11.2 and confirmation of the X-linked locus to Xq28. Genomics 34(1): 143-146.

Jackson MS, Rocchi M, Thompson G, Hearn T, Crosier M, Guy J, Kirk D, Mulligan L, Ricco A, Piccininni S, Marzella R, Viggiano L, Archidiacono N. (1999) Sequences flanking the centromere of human chromosome 10 are a complex patchwork of arm-specific sequences, stable duplications and unstable sequences with homologies to telomeric and other centromeric locations. Human Molecular Genetics 8(2): 205-215.

Jeffs P, Ashburner M. (1991) Processed pseudogenes in Drosophila. Proceedings of the Royal Society of London Biological Sciences 244(1310):151-159.

Jerome LA, Papaioannou VE. (2001) DiGeorge syndrome phenotype in mice mutant for the T-box gene, Tbx1. Nature Genetics 27: 286-291.

Johnson A, Minoshima S, Asakawa S, Shimizu N, Shizuya H, Roe BA, McDermid HE. (1999) A 1.5-Mb contig within the cat eye syndrome critical region at human chromosome 22q11.2. Genomics 57(2): 306-309.

Kawamura K, Shibata T, Saget O, Peel D, Bryant PJ. (1999) A new family of growth factors produced by the fat body and active on Drosophila imaginal disc cells. Development 126(2): 211-219.

Keegan LP, Gerber AP, Brindle J, Leemans R, Gallo A, Keller W, O'Connell MA. (2000) The properties of a tRNA-specific adenosine deaminase from Drosophila melanogaster support an evolutionary link between pre-mRNA editing and tRNA modification. Molecular and Cellular Biology 20(3): 825-833.

Kim J, Kim P, Hui CC. (2001) The VACTERL association: lessons from the Sonic hedgehog pathway. Clinical Genetics 59(5): 306-315.

Knoll JHM, Asamoah A, Pletcher BA, Wagstaff J. (1995) Interstitial duplication of proximal 22q phenotypic overlap with cat eye syndrome. American Journal of Medical Genetics 55: 221-224.

Kohlhase J, Wischermann A, Reichenbach H, Froster U, Engel W. (1998) Mutations in the SALL1 putative transcription factor gene cause Townes-Brocks syndrome. Nature Genetics 18(1): 81-83.

Langley CH, Montgomery E, Quattlebaum WF. (1982) Restriction map variation in the Adh region of Drosophila. Proceedings of the National Academy of Science USA 79(18): 5631-5635.

Leibovich SJ, Chen JF, Pinhal-Enfield G, Belem PC, Elson G, Rosania A, Ramanathan M, Montesinos C, Jacobson M, Schwarzschild MA, Fink JS, Cronstein B. (2002) Synergistic up-regulation of vascular endothelial growth factor expression in murine macrophages by adenosine A(2A) receptor agonists and endotoxin. American Journal of Pathology 160(6): 2231-2244.

Li S, Aksoy S. (2000) A family of genes with growth factor and adenosine deaminase similarity are preferentially expressed in the salivary glands of Glossina m. morsitans. Gene 252(1-2): 83-93.

Lindsay EA, Vitelli F, Su H, Morishima M, Huynh T, Pramparo T, Jurecic V, Ogunrinu G, Sutherland HF, Scambler PJ, Bradley A, Baldini A. (2001) Tbx1 haploinsufficieny in the DiGeorge syndrome region causes aortic arch defects in mice. Nature 410: 97-101.

Loftus BJ, Kim UJ, Sneddon VP, Kalush F, Brandon R, Fuhrmann J, Mason T, Crosby ML, Barnstead M, Cronin L, Deslattes Mays A, Cao Y, Xu RX, Kang HL, Mitchell S, Eichler EE, Harris PC, Venter JC, Adams MD. (1999) Genome duplications and other features in 12 Mb of DNA sequence from human chromosome 16p and 16q. Genomics 60(3): 295-308.

Long M, Langley CH. (1993) Natural selection and the origin of jingwei, a chimeric processed functional gene in Drosophila. Science 260(5104): 91-95.

Long M, Wang W, Zhang J. (1999) Origin of new genes and source for N-terminal domain of the chimerical gene, jingwei, in Drosophila. Gene 238(1): 135-141.

Lundin LG. (1993) Evolution of the vertebrate genome as reflected in paralogous chromosomal regions in man and the house mouse. Genomics 16(1): 1-19.

Maier SA, Podemski L, Graham SW, McDermid HE, Locke J. (2001) Characterization of the adenosine deaminase-related growth factor (ADGF) gene family in Drosophila. Gene 280(1-2): 27-36.

McDermid HE, Duncan AMV, Brasch JJA, Magenis HE, Sheely R, Burn J, Kardon N, Noel B, Schinzel A, Teshima I, White BN. (1986) Characterization of the supernumerary chromosome in cat eye syndrome. Science 232: 646-648.

McDermid HE, McTaggart KE, Riazi MA, Hudson TJ, Budarf ML, Emanuel BS, Bell CJ. (1996) Long-range mapping and construction of a YAC contig within the cat eyen syndrome critical region. Genome Research 6: 1149-1159.

McDermid HE, Morrow BE. (2002) Genomic disorders on 22q11. American Journal of Human Genetics 70(5): 1077-1088.

McTaggart KE, Budarf ML, Driscoll DA, Emanuel BS, Ferreira P, McDermid HE. (1998) Cat eye syndrome chromosome breakpoint clustering: identification of two intervals also associated with 22q11 deletion syndrome breakpoints. Cytogenetics and Cell Genetics 81: 222-228.

Mears AJ, Duncan AMV, Budarf ML, Emanuel BS, Sellinger B, Siegel-Bartlet J, Greenberg CR, McDermid HE. (1994) Molecular characterization of the marker chromosome associated with cat eye syndrome. American Journal of Human Genetics 55: 134-142.

Mears AJ, El-Shanti H, Murray JC, McDermid HE, Patil SR. (1995) Minute supernumerary ring chromosome 22 associated with cat eye syndrome: further delineation of the critical region. American Journal of Human Genetics 57: 667-673.

Mears AJ. PhD thesis (1995) Molecular characterization of duplications associated with cat eye syndrome. University of Alberta.

Michel PP, Marien M, Ruberg M, Colpaert F, Agid Y. (1999) Adenosine prevents the death of mesencephalic dopaminergic neurons by a mechanism that involves astrocytes. Journal of Neurochemistry 72(5): 2074-2082.

Murphy TD, Karpen GH. (1998) Centromeres take flight: alpha satellite and the quest for the human centromere. Cell 93(3): 317-320.

Netzer C, Rieger L, Brero A, Zhang CD, Hinzke M, Kohlhase J, Bohlander SK. (2001) SALL1, the gene mutated in Townes-Brocks syndrome, encodes a

transcriptional repressor which interacts with TRF1/PIN2 and localizes to pericentromeric heterochromatin. Human Molecular Genetics 10(26): 3017-3024.

Ohno S. (1970) *Evolution by gene duplication*. Springer Verlag, Berlin, Germany.

Pagon RA, Graham JM Jr, Zonana J, Yong SL. (1981) Coloboma, congenital heart disease, and choanal atresia with multiple anomalies: CHARGE association. Journal of Pediatrics 99(2): 223-227.

Regnier V, Meddeb M, Lecointre G, Richard F, Duverger A, Nguyen VC, Dutrillaux B, Bernheim A, Danglot G. (1997) Emergence and scattering of multiple neurofibromatosis (NF1)-related sequences during hominoid evolution suggest a process of pericentromeric interchromosomal transposition. Human Molecular Genetics 6(1): 9-16.

Reiss JA, Weleber RG, Brown MG, Bangs CD, Loverin EW, Magenis RE. (1985) Tandem duplication of proximal 22q: a cause of cat eye syndrome. American Journal of Medical Genetics 20: 165-171.

Riazi MA. PhD thesis (1998) Transcriptional mapping in the proximal region of human chromosome 22. University of Alberta.

Riazi MA, Brinkman-Mills P, Nguyen T, Pan H, Phan S, Ying F, Roe BA, Tochigi J, Shimizu Y, Minoshima S, Shimizu N, Buchwald M, McDermid HE. (2000) The human homolog of insect-derived growth factor, CECR1, is a candidate gene for features of cat eye syndrome. Genomics 64(3): 277-285.

Riazi MA, Buchwald M. (2001) Transgenic expression of *CECR1* results in developmental abnormalities similar to features of cat eye syndrome. American Journal of Human Genetics 69: 635.

Ried K, Rao E, Schiebel K, Rappold GA. (1998) Gene duplications as a recurrent theme in the evolution of the human pseudoautosomal region 1: isolation of the gene ASMTL. Human Molecular Genetics 7(11): 1771-1778.

Rivkees SA, Zhao Z, Porter G, Turner C. (2001) Influences of adenosine on the fetus and newborn. Molecular Genetics of Metabolism 74(1-2): 160-171.

Rogalla P, Lemke I, Kazmierczak B, Bullerdiek J. (2000) An identical HMGIC-LPP fusion transcript is consistently expressed in pulmonary chondroid hamartomas with t(3;12)(q27-28;q14-15). Genes Chromosomes Cancer 29(4): 363-366.

Schachenmann G, Schmid W, Fraccaro M, Mannini A, Tiepolo L, Perona GP, Sartori E. (1965) Chromosomes in coloboma and anal atresia. Lancet 2: 290.

Sambrook J, Russell D. (2001) *Molecular Cloning (A Laboratory Manual)*, Cold Spring Harbour Laboratory Press, Cold Spring Harbour, New York.

Schinzel A, Schmid W, Fraccaro M, Tiepolo L, Zuffardi O, Opitz JM, Lindsten J, Zetterquist P, Enell H, *et al.* (1981) The "cat-eye-syndrome": Dicentric small marker

chromosome probably derived from a No. 22 (tetrasomy 22 pter-q11) associated with a characteristic phenotype. Human Genetics 57: 148-158.

Schueler MG, Higgins AW, Rudd MK, Gustashaw K, Willard HF. (2001) Genomic and genetic definition of a functional human centromere. Science 294(5540): 109-115.

Smith GP. (1976) Evolution of repeated DNA sequences by unequal crossover. Science 191: 528-535.

Smit AFA, Green P. Unpublished data.

Swillen A, Vogels A, Devriendt K, Fryns JP. (2000) Chromosome 22q11 deletion syndrome: Update and review of the clinical features, cognitive-behavioural spectrum, and psychiatric complications. American Journal of Medical Genetics 97(2): 128-135.

Strong WB. (1968). Familial syndrome of right-sided aortic arch, mental deficiency, and facial dimorphism. Journal of Pediatrics 73: 882-888.

Tellier AL, Amiel J, Delezoide AL, Audollent S, Auge J, Esnault D, Encha-Razavi F, Munnich A, Lyonnet S, Vekemans M, Attie-Bitach T. (2000) Expression of the PAX2 gene in human embryos and exclusion in the CHARGE syndrome. American Journal of Human Genetics 93(2): 85-88.

Trask BJ, Massa H, Brand-Arpon V, Chan K, Friedman C, Nguyen OT, Eichler E, van den Engh G, Rouquier S, Shizuya H, Giorgi D. (1998) Large multi-chromosomal duplications encompass many members of the olfactory receptor gene family in the human genome. Human Molecular Genetics 7(13): 2007-2020.

Williamson JR. (1994) G-quartet structures in telomeric DNA. Annual Review of Biophysical and Biomolecular Structures 23: 703-730.

Winston JH, Hanten GR, Overbeek PA, Kellems RE. (1992) 5' flanking sequences of the murine adenosine deaminase gene direct expression of a reporter gene to specific prenatal and postnatal tissues in transgenic mice. Journal of Biological Chemistry 267(19): 13472-13479.

Xu PA, Kellems RE. (2000) Function of murine adenosine deaminase in the gastrointestinal tract. Biochemical and Biophysical Research Communications 269(3): 749-757.

Yeung CY, Ingolia DE, Roth DB, Shoemaker C, Al-Ubaidi MR, Yen JY, Ching C, Bobonis C, Kaufman RJ, Kellems RE. (1985) Identification of functional murine adenosine deaminase cDNA clones by complementation in Escherichia coli. Journal of Biological Chemistry 260(18): 10299-10307.

Zurovec M, Dolezal T, Gazi M, Pavlova E, Bryant PJ. (2002) Adenosine deaminase-related growth factors stimulate cell proliferation in Drosophila by depleting

extracellular adenosine. Proceedings of the National Academy of Science of the USA 99(7): 4403-4408.